Online optimization methods and application for accelerators

Seminar at MSU/FRIB

9/23/2022

Xiaobiao Huang

SLAC National Accelerator Laboratory

U.S. DEPARTMENT OF ENERGY

Office of Science

SLAC NATIONAL ACCELERATOR LABORATORY

Outline

- Overview of automated tuning
- Methods for automated tuning
  - Deterministic algorithms: simplex, RCDS
  - Stochastic algorithms: GA, PSO, MG-GPO
- Application examples
  - Storage ring vertical emittance
  - Linac transmission
  - Storage ring nonlinear dynamics
- Discussion
- Summary

Overview of automated tuning

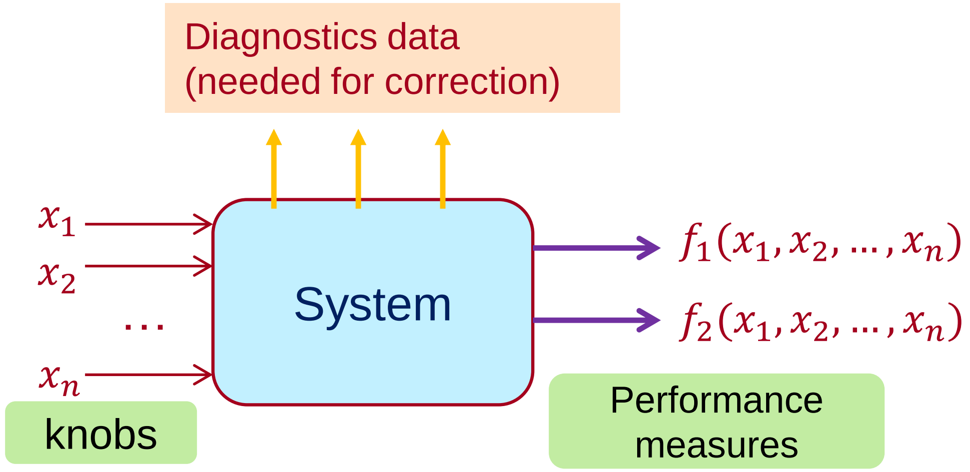
- Accelerators usually do not behave exactly as models predict at commissioning
  - Many error sources, both in machine and in model
- Accelerator physicists have always relied on tuning, i.e., online optimization, to improve machine performance
  - But tuning by hand is limited in speed, scale, and complexity
- Effort to automate online tuning goes back a long way
  - SLC — N.J. Walker et al, PAC'93 (1993)
  - KEKB — J.W. Flanagan et al, ICAPC (1998)
  - APS — L. Emery et al, PAC2003
  - …

Beam-based correction vs. optimization

- Beam-based methods to turn knobs may be grouped into two camps: correction and optimization
- **Beam-based correction**: to deduce the required knob changes toward a machine state directly from beam measurements
  - w/ sufficient diagnostics to determine the machine state (orbit, optics, coupling, etc) in a deterministic fashion
  - w/ sufficient understanding of the system to relate machine state to knob changes (e.g., through response matrix)
  - Example: orbit correction, LOCO, tune/chromaticity correction
- **Beam-based optimization**: to turn knobs to minimize or maximize the performance metric(s)
  - Machine is considered a black box
  - Although, machine state measurements and knowledge of the system could be used to improve optimization efficiency (depending on algorithms)

Why beam-based optimization?

- Sometimes the correction approach can’t be done
  - No or insufficient diagnostics to determine the machine state
  - The correction target (in terms of machine state) is not determined
  - Difficult to invert the problem (i.e., to go from machine state to knob changes)
  - Not enough knobs for correction
- Sometimes the optimization approach can yield better performance



Methods for automated tuning

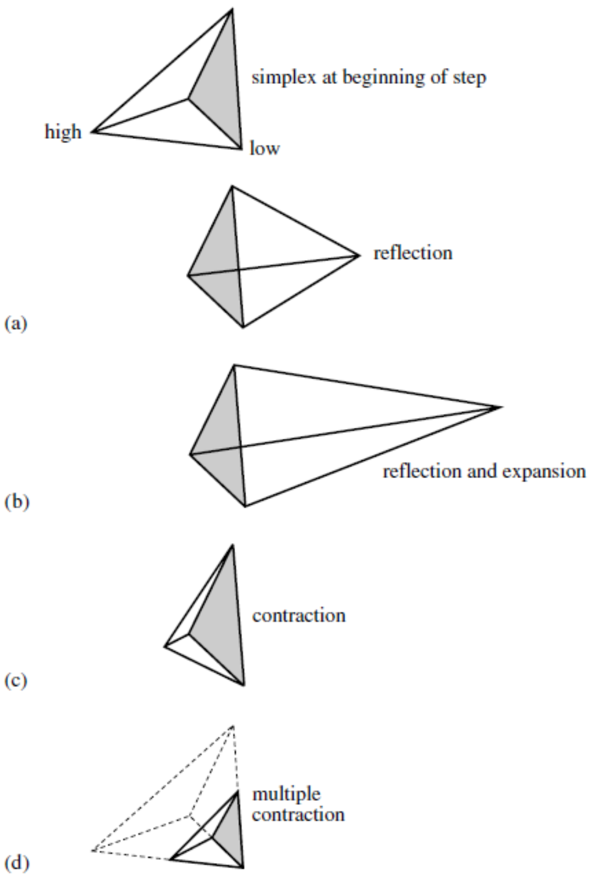
- Optimization algorithms w/ special emphasis:
  - Resistance to noise
  - High efficiency
  - Robustness (works most of the time)
- Deterministic methods
  - Grid scan, simplex, …
  - Gradient based methods
  - Robust conjugate direction search (RCDS)
- Stochastic methods
  - Random search, genetic algorithms, particle swarm, …
  - Machine learning based algorithms
  - Multi-generation Gaussian process optimizer (MG-GPO)

There are so many methods out there. I can only focus on my favorites in this talk.

Nelder-Mead Simplex

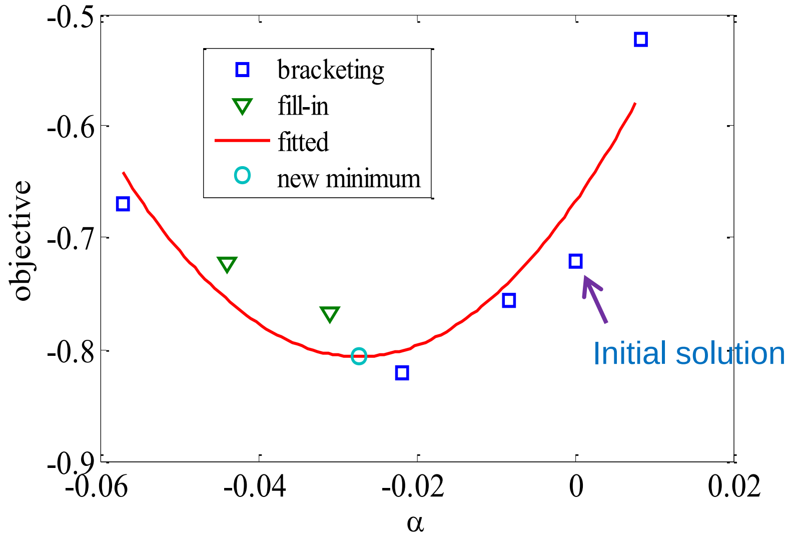
- The simplex method is one of the most popular gradient-free multi-variable optimization methods
  - Powerful, easy to use, …
  - A direct search method, which relies on comparison of function values to make decisions on 5 operations
  - Mostly search along the line connecting the worst vertex and the center-of-mass point on the opposing face

    - It is sensitive to measurement noise
    - Initial simplex size is important for online application
    - May converge to local minimum



The RCDS method

- It was developed specifically to resist the effect of noise
- It employs iterative 1-D optimizations (preferably along conjugate directions), with the robust 1-D optimizer



For robust 1D optimizer

- Bracket minimum w/ consideration of noise level
- Find minimum w/ parabolic fitting

X. Huang et al, Nucl. Instr. Methods, A 726 (2013) 77-83.

The RCDS method has found applications on many accelerators.

Genetic algorithms and particle swarm optimizer

- Both are population-based, evolutionary algorithms
  - Generate new solutions through stochastic operations
  - Use non-dominated sorting to select survivors

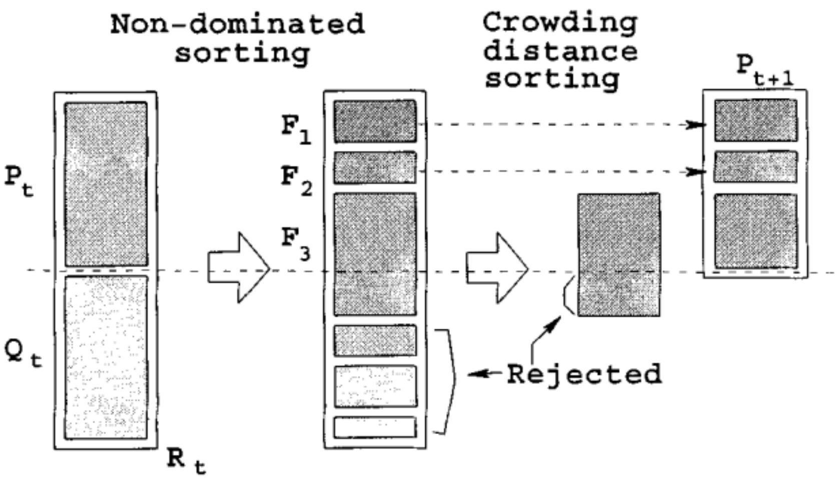


Fig. 2. NSGA-II procedure.

K. Deb, IEEE Transactions On Evolutionary Computation Vol 6, No 2,April 2002

PSO, J. Kennedy, Proceedings of ICNN'95, 1995

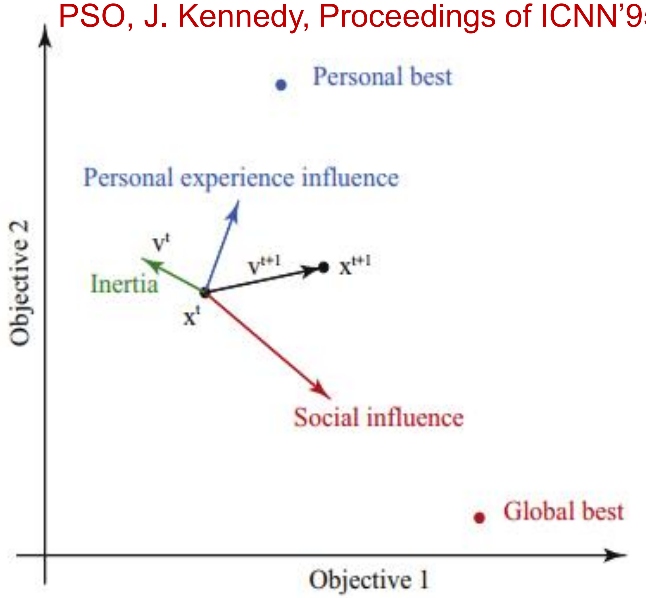


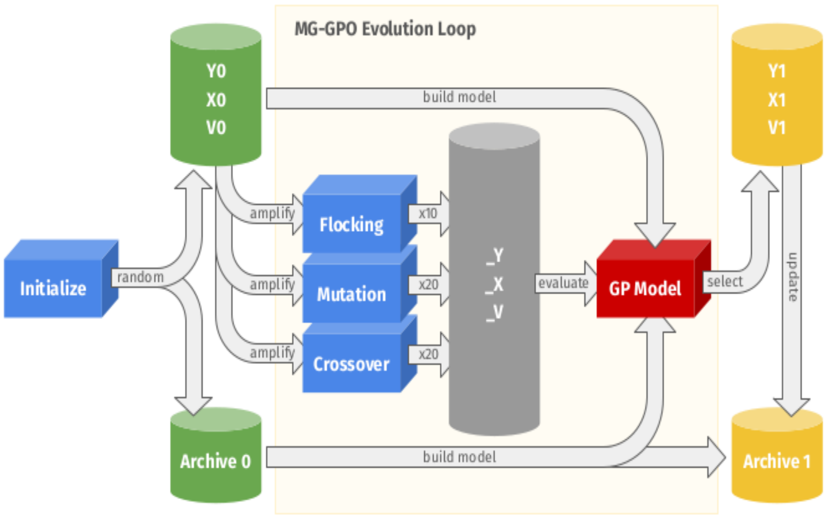
Fig. 1. Velocity and position updates in PSO.

Illustration from X. Pang, et al NIMA 2014

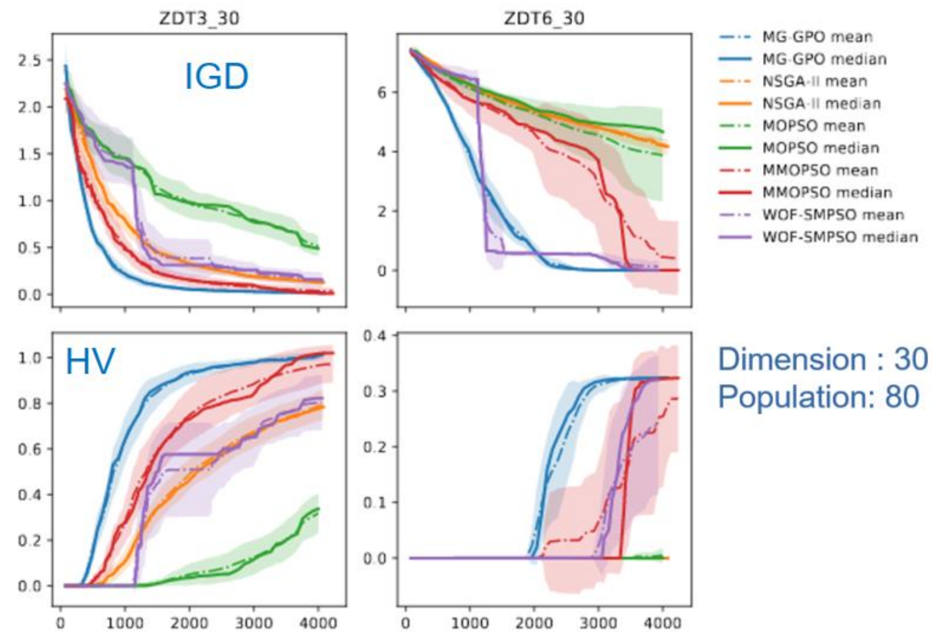
The MG-GPO method

- It is a multi-objective, stochastic optimization algorithm
- Similar to GA, PSO, etc., but assisted with ML to improve efficiency

It uses GP models to filter for trial solutions



X. Huang, M. Song, Z. Zhang, arXiv 1907.00250 (2019)
Diagram from Z. Zhang, M. Song, X. Huang, ML Science and Technology, 2, 015014 (2021)

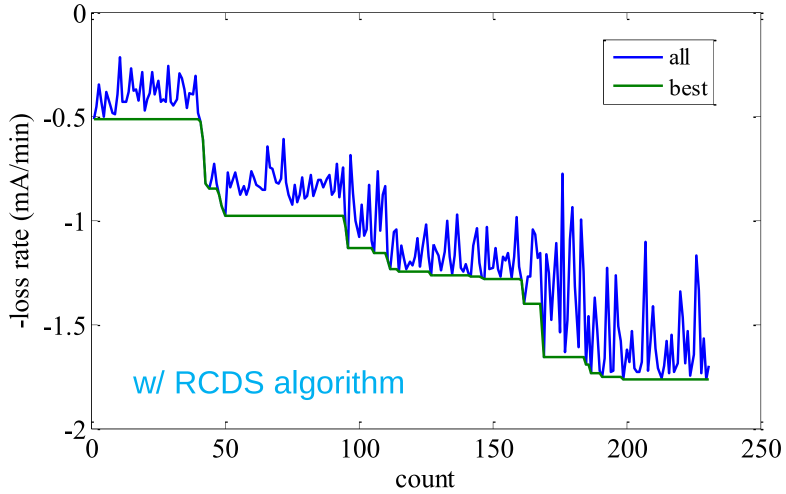


Simulation tests to benchmark MG-GPO performance
IGD: inverted generational distance
HV: hyper volume

Application example: minimization of storage ring vertical emittance for SPEAR3

- Vertical emittance in a storage ring comes from linear x-y coupling and vertical dispersion
  - Both can be compensated with skew quadrupoles
- Objective: vertical beam size, or indirectly, Touschek beam loss rate
- Knobs: 13 skew quadrupoles

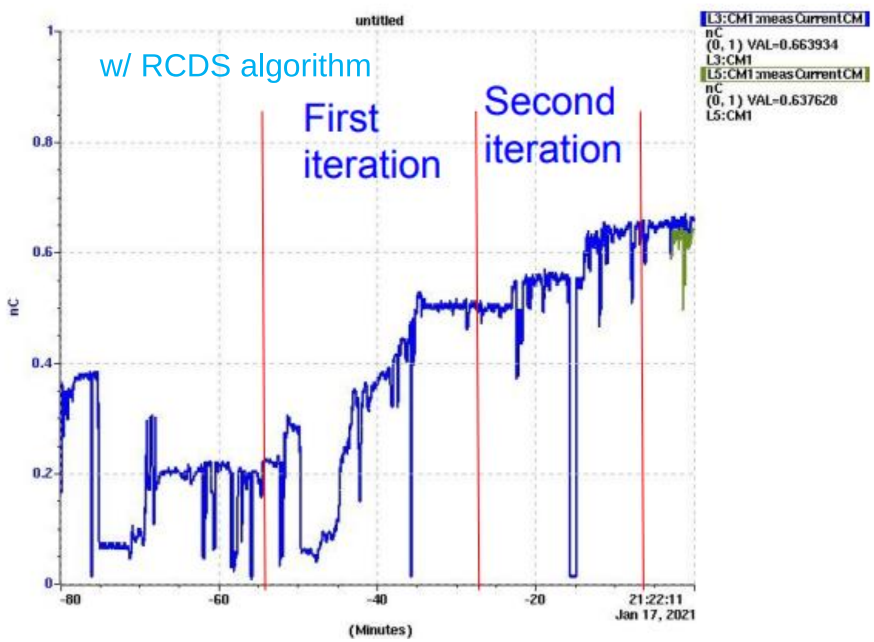
Using loss rate (normalized) as objective

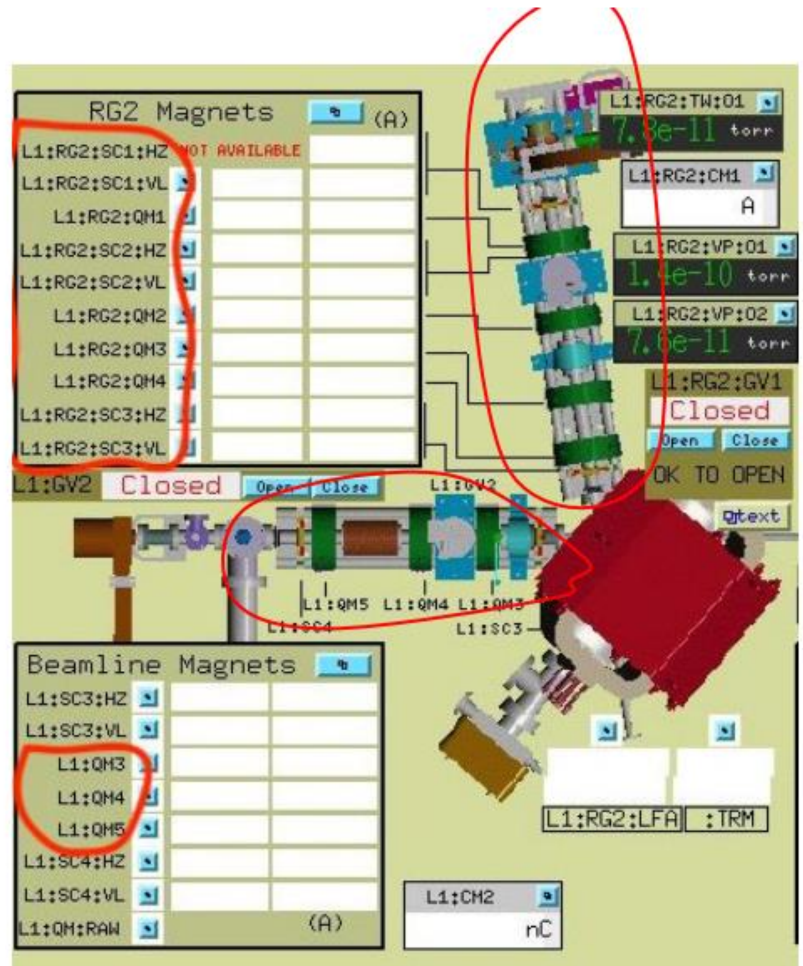


- Start with all skew quadrupoles off
- Knob values for the best solution are similar to LOCO (fitting orbit response matrix) solution
- Reaches lower coupling ratio than LOCO

Application example: Linac front end for APS

- To maximize transmission of beam at linac front end
- Objective: beam charge in linac
- Knobs: 12 magnets (corrector and quads)



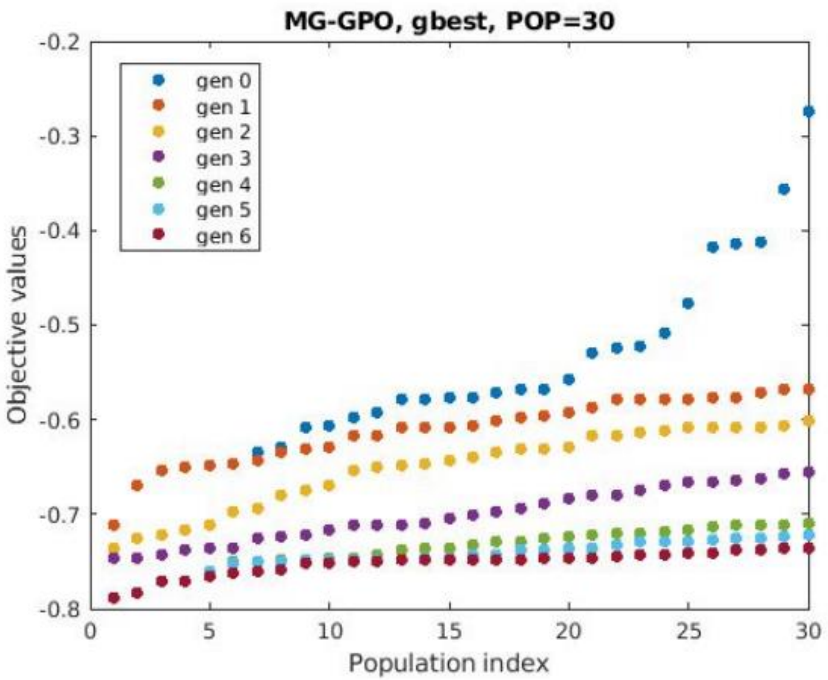


H. Shang, et al, IPAC'21 (2021)

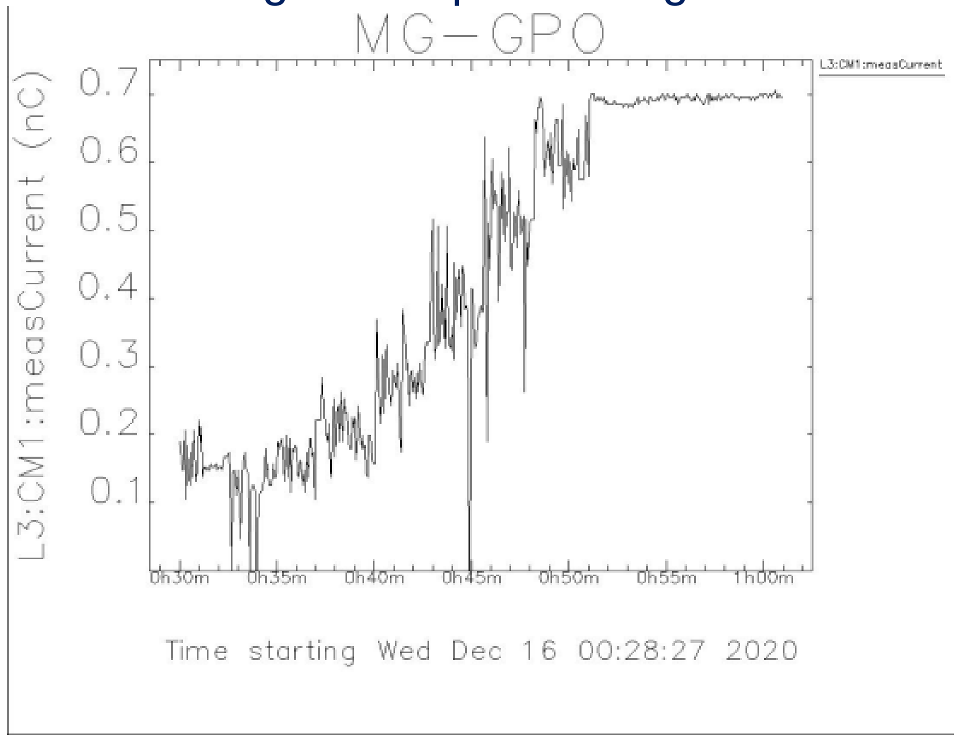
Same setup w/ MG-GPO

- MG-GPO is less sensitive to starting points

Starting from operation condition



Starting from a poor configuration



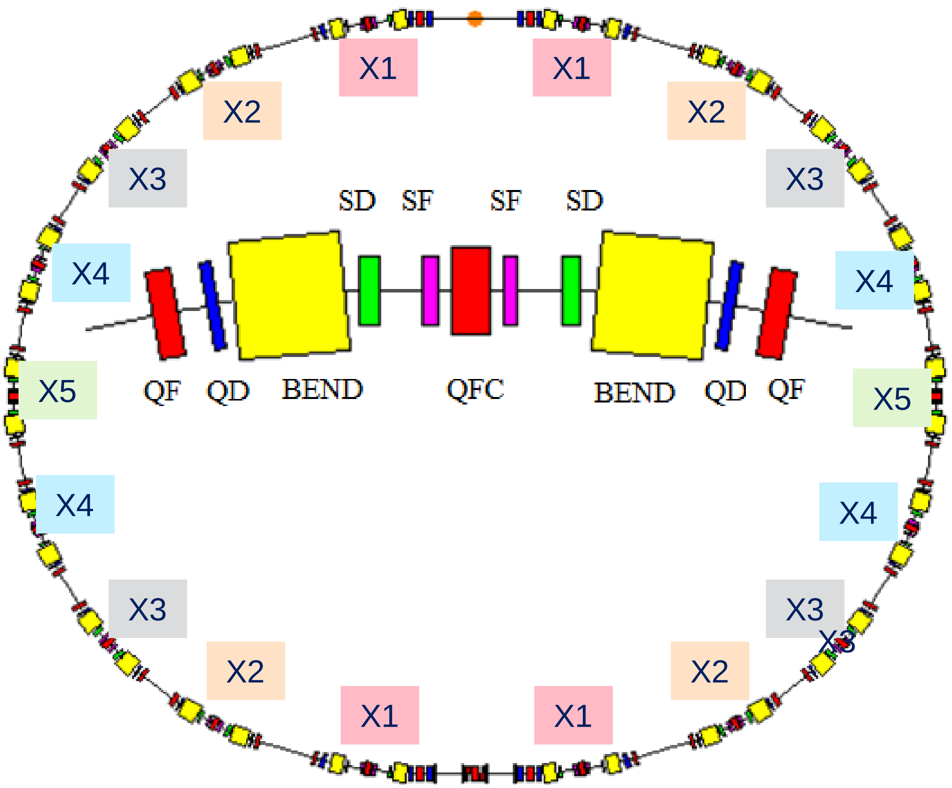
H. Shang, et al, IPAC'21 (2021)

Online optimization of storage ring nonlinear dynamics

- Problem description: to maximize the dynamic aperture (DA) and momentum acceptance (MA) of a storage ring with linear or nonlinear magnets
- A classic example where beam-based optimization is needed
  - No sufficient diagnostics to establish machine state w/ good DA/MA
  - No target machine state for optimal DA/MA
  - Not enough knobs to correct the potentially relevant machine state variables
  - But DA and MA both can be directly measured or represented w/ proxies
- As newer storage rings push for ultra low emittances, nonlinear beam dynamics has become a dominant factor in design considerations.
  - Online optimization provides a path to realize the design DA/MA performance

The SPEAR3 DA optimization setup – knobs

- SPEAR3 is a third generation light source at SLAC.
  - The lattice consists of 18 DBA cells
  - There are a total of 10 sextupole families.

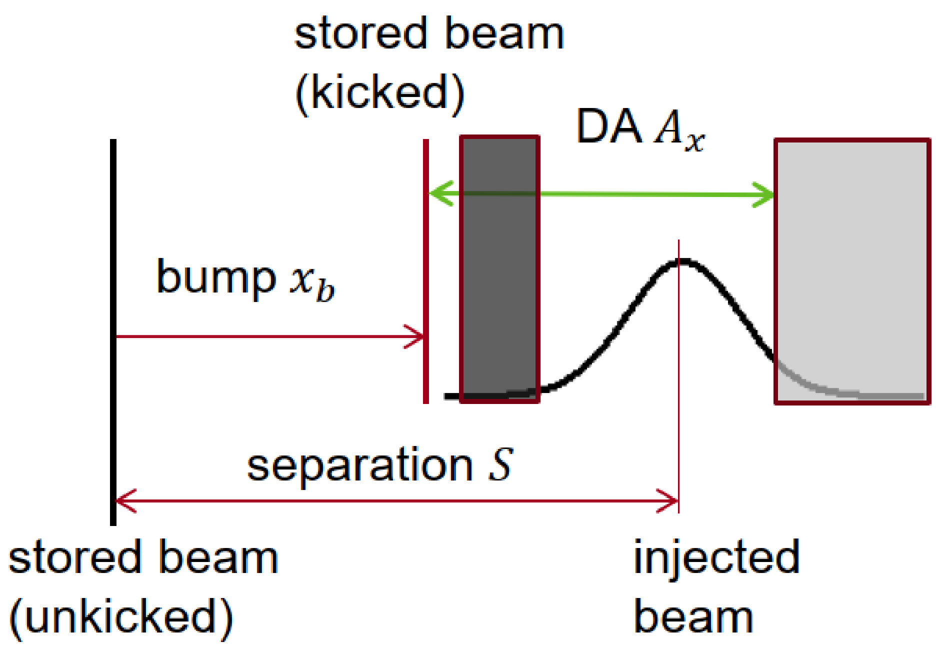


Parameter	value
Energy	3 GeV
Circumference	234 m
Emittance (ϵ_x)	10 nm
Tunes (ν_x, ν_y)	14.106, 6.177

- All sextupoles are in dispersive region.
- Form 8 comb-knobs that do not change chromaticities w/ response matrix

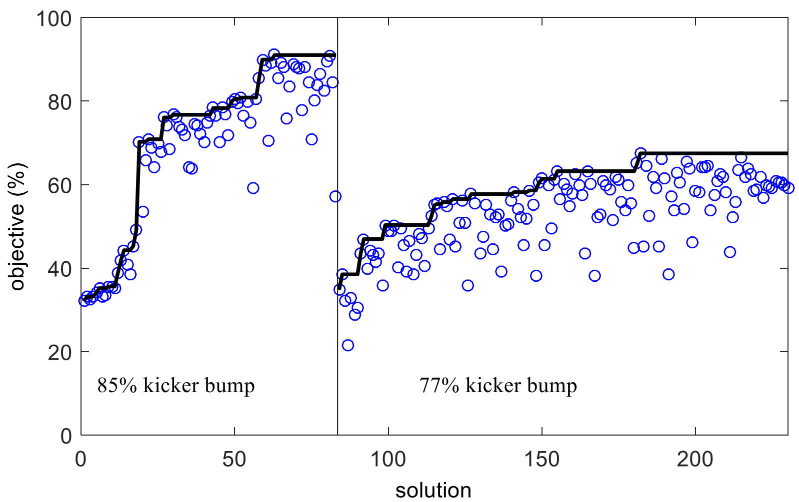
Objective – SPEAR3 DA

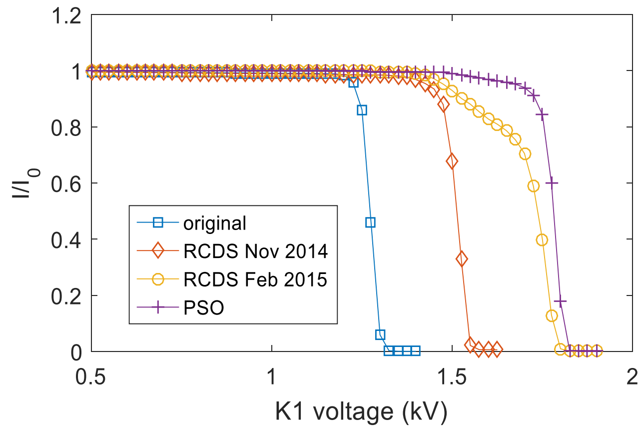
- The injection efficiency is used as a proxy for DA
  - Direct measurement of DA is time consuming
  - Reduce the kicker bump so that injection efficiency is sensitive to DA changes
  - Decrease injector beam intensity and tune for machine stability
  - Average over a 10-second period to reduce measurement noise



Optimization results – SPEAR3 DA

- Optimization led to substantial enlargement of DA for SPEAR3
  - DA went from 15 mm to 20 mm
  - No decrease of MA (confirmed w/ RF voltage scan)



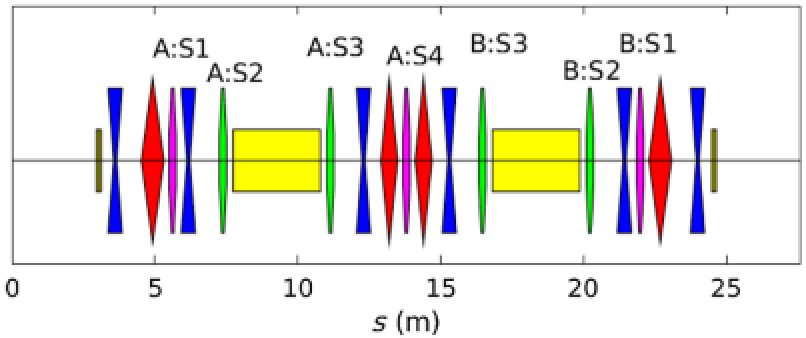


X. Huang, J. Safranek, PRSTAB 18, 084001 (2015)

DA optimization has been critical to implementation of lower emittance lattices for SPEAR3.

DA/MA optimization for Advanced Photon Source (APS)

- APS: a 7-GeV, 1104 m storage ring w/ 40 DBA cells
- Normally DA is limited by physical aperture at an ID for APS
  - Enlarged physical acceptance by decreasing horizontal beta function at the ID location (β_x from 20 m to 15 m, then to 10 m)
- Tuning knobs
  - There are 280 sextupoles, all individually powered
  - Formed 7 sextupole families, assuming a 20-fold periodicity and symmetry about ID straights



Use chromaticity matrix to obtain 5 free knobs that do not change chromatcities

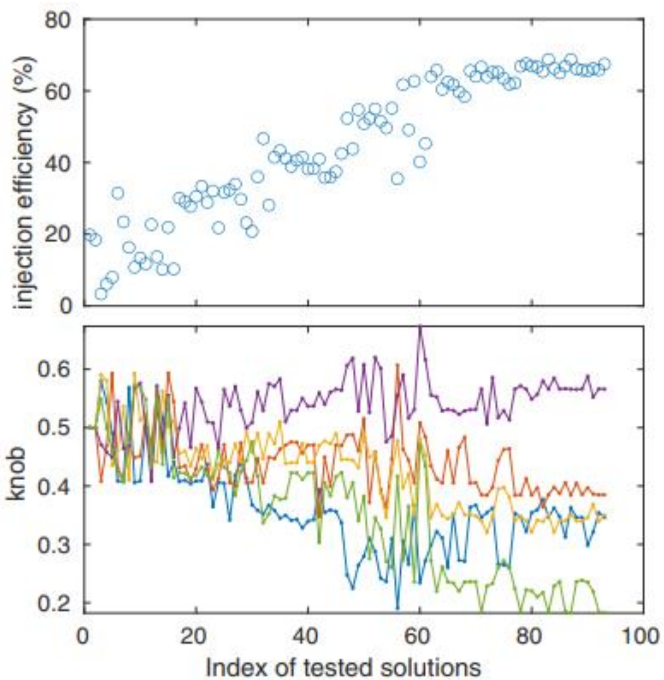
Objectives – APS DA/MA optimization

- Injection efficiency was used as proxy for DA
  - Decreasing injection efficiency with kicker bump mismatch
  - Measure injection efficiency for 2 seconds
  - Dump and refill for every measurement
- Beam lifetime was used as proxy for MA
  - 24 mA in 6 bunches (Touschek loss dominates)
  - Monitor for 20 seconds
  - Normalize lifetime w/ beam current (squared) and vertical beam size (measured by pinhole camera)

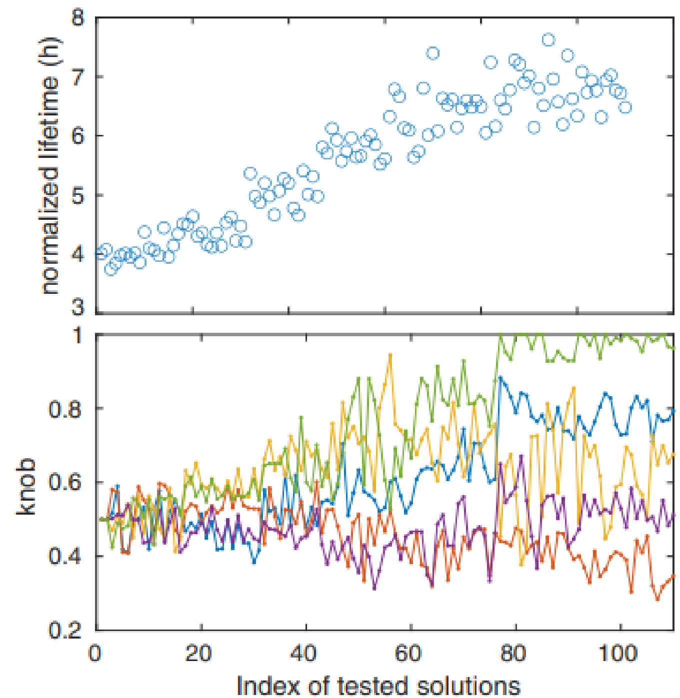
Error sigma is about 1% for injection efficiency and 2% for lifetime.

APS optimization results – single objective

- Single objective optimization for DA or MA worked well
  - But best solution for DA does not have good MA and vice versa
  - An alternate (DA/MA), iterative approach did not work



APS DA optimization with MG-GPO

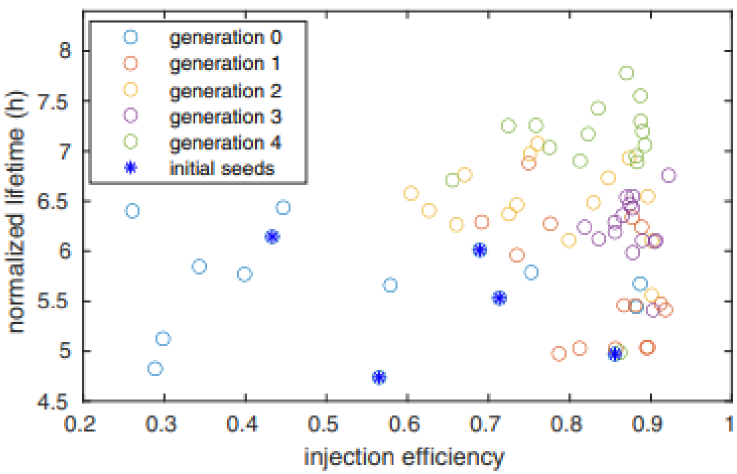


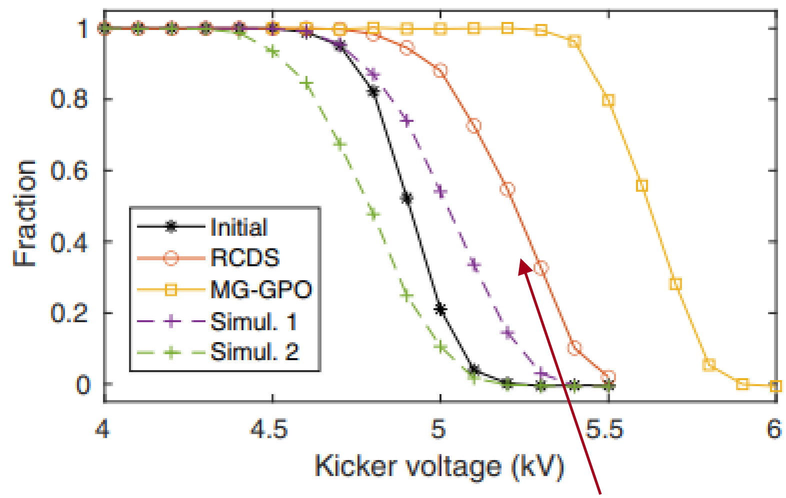
APS MA optimization with MG-GPO

L. Emery, H. Shang, Y. Sun, X. Huang, PRAB 24, 082802 (2021)

Simultaneous DA and MA optimization at APS

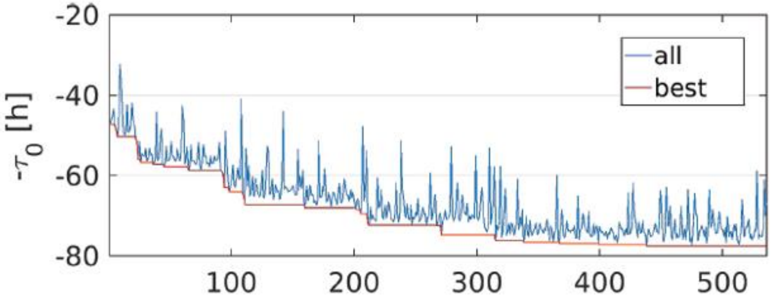
- Start w/ a population that consists of best solutions from DA and MA optimizations.
- Evaluate DA and MA separately to reduce total beam loss
  - Better than evaluate and dump beam for every solution



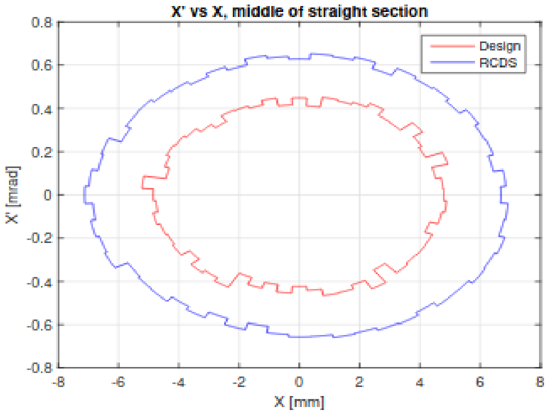


- Note the RCDS solution reached the limit by physical acceptance ($\beta_x =$ 15 m), while MG-GPO solution is with $\beta_x = 10$ m.
- Solutions from simulation did not correspond to large measured DA

DA or MA optimization on other storage rings



Optimization of ESRF Touschek lifetime
S. M. Liuzzo, et al, IPAC'16



Dynamic aperture optimization
for MAX-IV
D. K. Olsson, IPAC'18

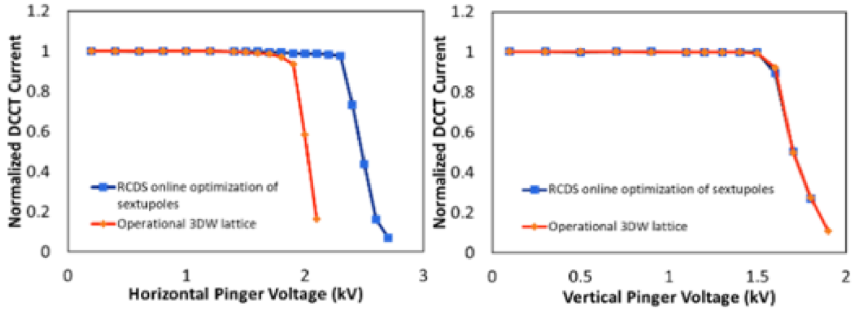


Figure 2: Horizontal (left) and vertical (right) DA before (red) and after (blue) sextupole optimization in the 3DW lattice.

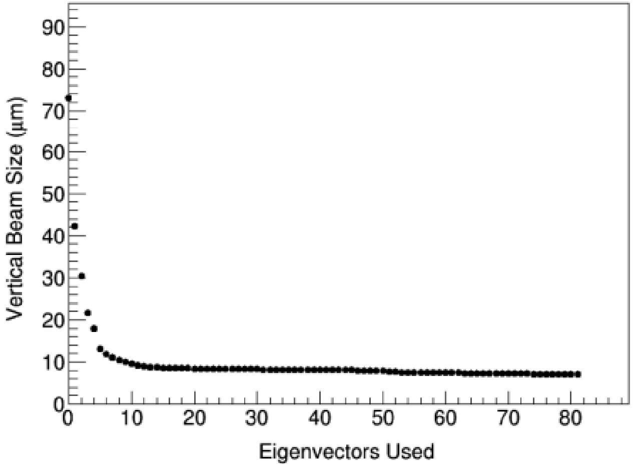
Dynamic aperture optimization for
NSLS-II
X. Yang, IPAC'22

It seems typical to get ~20% increase of DA w/ online optimization.

Discussion on application considerations

- Measurement noise
  - It is preferable to reduce noise by increasing measurement time or with averaging
- Forming conjugate directions
  - Knobs independent to each other in terms of impact to objective(s)
  - May use models for calculation
- Reduction of dimension
  - Algorithms typically work best w/ up to 10-15 knobs
  - More knobs needs more evaluations to converge

Can apply SVD to model-calculated Hessian to reduce dimension



Vertical beam size minimization on CESR w/ dimension reduction
W. Bergan, et al, PRAB 22, 054601 (2019)

Summary

- Automated tuning has become common in the arsenal of control-room accelerator physicists
  - Has advantages over beam-based correction in some ways
- There have been a number of online optimization algorithms that have been widely tested with real-life problems
- Online optimization have been successful for many important applications
  - Storage ring nonlinear dynamics is a topic of special importance
- Further development in online optimization is expected