Constrained, Multi-objective, and Parameterized Optimization

David Bindel

30 June 2022

Department of Computer Science
Cornell University

What Makes a Good Stellarator (2019 edition)?

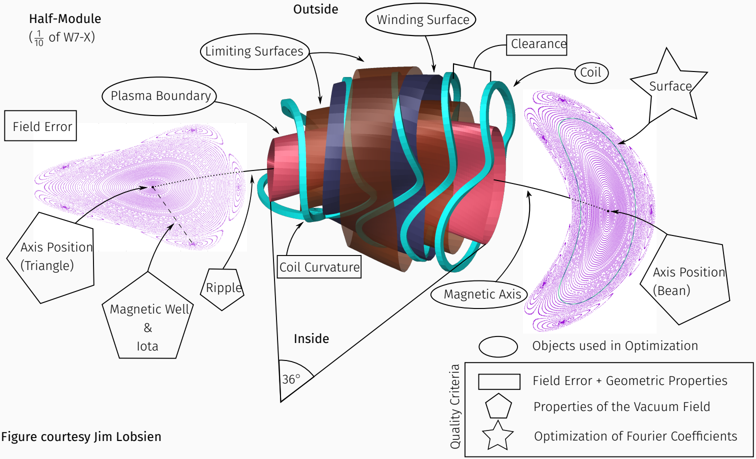


Figure courtesy Jim Lobsien

Challenges (2019 edition)

1. Costly and "black box" physics computations
    - Each step: MHD equilibrium solve, transport, coil design, ...
    - Several times per step for finite-difference gradients
2. Managing tradeoffs
    - How do we choose the weights in the χ^2 measure? By gut?
    - Varying the weights does not expose tradeoffs sensibly
3. Dealing with uncertainties
    - What you simulate $\neq$ what you build!
4. Global search
    - How to avoid getting stuck in local minima?

Progress

- Collaboration has made a lot of progress on
  - Faster simulations, with derivatives
  - Optimizing under uncertainty
- Limited progress on global search (TuRBO)
- Still less on tradeoffs and constraints

Background: Unconstrained Optimization

Assume $\phi : \mathbb{R}^n \to \mathbb{R}$ is $\mathcal{C}^2$, seek

$$\text{minimize } \phi(x) \text{ over } x \in \mathbb{R}^n$$

Standard (local) strategy from an adequate guess x^0:

- Approximate ϕ near x^k by a model (usu. quadratic)
- Minimize the model to find x^{k+1} (linear algebra)
- Avoid over-stepping by line search, trust region, etc (globalization)

Lots of room for cleverness, using problem structure.

Newton Framework

Quadratic model:

$$\phi(x^k + u) \approx \phi(x^k) + \nabla\phi(x^k)^T u + \frac{1}{2} u^T H_\phi(x^k) u$$

Model gradient: $\nabla\phi(x^k) + H_\phi(x^k)u$.
Minimized at $u = -H_\phi(x^k)^{-1}\nabla\phi(x^k)$ (if H pos def).

Lots of standard methods fudge H in some way:

- For convergence (e.g. trust region)
- For cost and convenience (e.g. BFGS)

Quadratic convergence $\implies$ asymptotically get Newton steps.

Nonlinear Least Squares

$$\phi(x) = \frac{1}{2}\|f(x)\|^2 \text{ where } f : \mathbb{R}^n \to \mathbb{R}^m; \quad \nabla\phi(x) = J(x)^T f(x), J(x) = f'(x)$$

Gauss-Newton idea:

$$\text{minimize } \|f(x^k) + J(x^k)p^k\|^2$$

and set $x^{k+1} = x^k + \alpha_k p^k$. Modified Newton with

$$H_\phi(x) = J(x)^T J(x) + \sum_{k=1}^{m} f_k(x) H_{\phi_k}(x) \approx J(x)^T J(x).$$

Levenberg-Marquardt: regularize Gauss-Newton

$$\text{minimize } \|f(x^k) + J(x^k)p^k\|^2 + \lambda_k^2 \|D_k x^k\|^2$$

where often $D_k = I$ (Levenberg) or $D_k^2 = \textbf{diag}\, J^T J$ (Marquardt).
Hessian $\approx J(x_k)^T J(x_k) + \lambda_k^2 D_k^2$.

Nonlinear LS Convergence

Gauss-Newton and Levenberg-Marquardt:

- Quadratic convergence when $f(x^*) = 0$, otherwise linear
- Linear rate depends on conditioning of $\kappa(J)$, $\|J'\|$, $\|f(x^*)\|$, and regularization or step size

A Common Approach

Put everything we care about in a nonlinear LS problem

- $f_k(x)$ is deviation from kth target
- Add some weighting (chosen by the user)

But is this actually what we want?

- Choice of target values is unclear
- Choice of weights is unclear

And there are reasons for numerical nervousness:

- Maybe too few objectives (underdetermined LS problems)
- Maybe poorly conditioned (esp. with “large” weights)
- May not have small residual

Tackling Constraints

General problem

$$\text{minimize } \phi(x) \text{ s.t. } \begin{cases} c_j(x) = 0, & j \in \mathcal{E} \\ c_j(x) \leq 0, & j \in \mathcal{I} \end{cases}$$

Convert into unconstrained optimization / nonlinear equation solving problem with:

- Fewer degrees of freedom (constraint elimination)
- Same degrees of freedom (penalties and barriers)
- More degrees of freedom (Lagrange multipliers)

Constraint elimination usually only for linear constraints.

KKT Conditions

$$\text{minimize } \phi(x) \text{ s.t. } \begin{cases} c_j(x) = 0, & j \in \mathcal{E} \\ c_j(x) \leq 0, & j \in \mathcal{I} \end{cases}$$

Define the Lagrangian

$$L(x, \lambda, \mu) = \phi(x) + \sum_{i \in \mathcal{E}} \lambda_i c_i(x) + \sum_{i \in \mathcal{I}} \mu_i c_i(x).$$

KKT conditions are

$$\begin{aligned} \nabla_x L(x^*) &= 0 && \\ c_i(x^*) &= 0, \quad i \in \mathcal{E} && \text{equality constraints} \\ c_i(x^*) &\leq 0, \quad i \in \mathcal{I} && \text{inequality constraints} \\ \mu_i &\geq 0, \quad i \in \mathcal{I} && \text{non-negativity of multipliers} \\ c_i(x^*)\mu_i &= 0, \quad i \in \mathcal{I} && \text{complementary slackness} \end{aligned}$$

Penalties and Barriers

Want to minimize

$$\text{minimize } \phi(x) \text{ s.t. } \begin{cases} c_j(x) = 0, & j \in \mathcal{E} \\ c_j(x) \leq 0, & j \in \mathcal{I} \end{cases}$$

Instead minimize for small γ

$$\psi_\gamma(x) = \phi(x) + \frac{1}{2\gamma} \sum_{i \in \mathcal{E}} c_i(x)^2 - \gamma \sum_{i \in \mathcal{I}} \log(-c_i(x)).$$

Note that at minimizer x^*:

$$\nabla \psi_\gamma(x^*) = \nabla \phi(x^*) + \sum_{i \in \mathcal{E}} \tilde{\lambda}_i \nabla c_i(x^*) + \sum_{i \in \mathcal{I}} \tilde{\mu}_i \nabla c_i(x^*)$$

where Lagrange multiplier estimates come from the c_i:

$$\tilde{\lambda}_i = c_i(x^*)/\gamma, \quad \tilde{\mu}_i = \gamma/c_i(x^*)$$

Standard trick: Penalty to estimate multipliers.

Managing Tradeoffs

What about using nonlinear least squares for tradeoffs?

More generally, consider $f : \mathbb{R}^n \to \mathbb{R}^m$, maybe minimize

$$w^T f(x) = \sum_{k=1}^{m} w_k f_k(x).$$

Incompleteness of χ-square Combination

Structural Optimization 14, 63–69 © Springer-Verlag 1997

A closer look at drawbacks of minimizing weighted sums of objectives for Pareto set generation in multicriteria optimization problems

I. Das and J.E. Dennis

Department of Computational and Applied Mathematics, Rice University of Houston, TX 77251-1892, USA

June 4, 2015 Matt Landreman

Some optimal solutions to a smooth multi-objective problem cannot be found by minimizing a total χ^2

Exploring the Pareto Frontier

x **dominates** y if

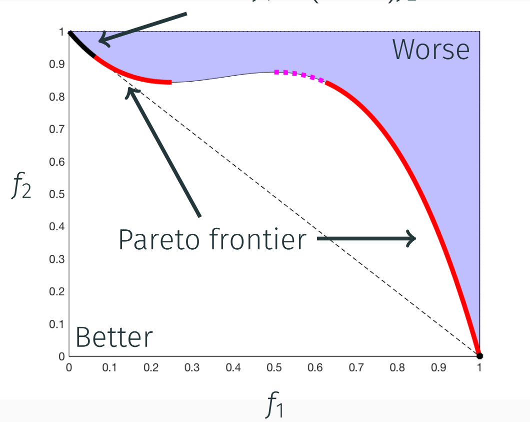
$$\forall k, f_k(x) \leq f_k(y)$$

and not all strict.

Best points are:
Pareto optimal,
aka non-dominated,
aka non-inferior,
aka non-efficient.

Form **Pareto frontier**

Minimize $\alpha f_1 + (1-\alpha) f_2$



Minimizing $\sum_k \alpha_k f_k$ only explores convex hull!
Other methods sample / approximate the full frontier.

First-order condition

Stationary condition:

$$\{J(x)u : u \geq 0\} \cap \mathbb{R}^n_+ = \{0\}.$$

Fritz John stationary condition: for some $\lambda \geq 0, \lambda \neq 0$

$$J(x)^T \lambda = 0.$$

Follows via Motzkin's theorem of the alternative: if A and C are given matrices, can either solve

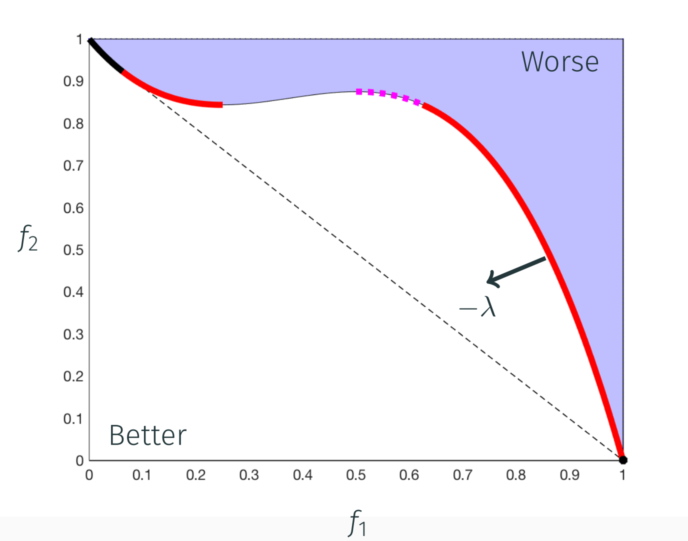
$$Ax < 0, \quad Cx \leq 0$$

or

$$A^T \lambda + C^T \mu = 0, \quad \lambda \geq 0, \lambda \neq 0, \mu \geq 0$$

But not both.

Fritz John multiplier geometry



Fritz John vs KKT

Fritz John condition (with constraints): Weak Pareto for

$$\text{minimize } f(x) \text{ s.t. } c(x) \leq 0$$

requires $\lambda \geq 0$ and $\mu \geq 0$ not both all zero such that

$$\begin{aligned} \lambda^T f'({}^*x) + \mu^T c'(x^*) &= 0 \\ \mu_i c_i(x^*) &= 0 \end{aligned}$$

Very similar to KKT conditions for constrained opt:

$$\begin{aligned} \nabla_x L(x^*) &= 0, & & L(x, \lambda, \mu) = \phi(x) + \lambda^T c_{\mathcal{E}}(x) + \mu^T c_{\mathcal{I}}(x) \\ c_i(x^*) &= 0, \quad i \in \mathcal{E} & & \text{equality constraints} \\ c_i(x^*) &\leq 0, \quad i \in \mathcal{I} & & \text{inequality constraints} \\ \mu_i &\geq 0, \quad i \in \mathcal{I} & & \text{non-negativity of multipliers} \\ c_i(x^*)\mu_i &= 0, \quad i \in \mathcal{I} & & \text{complementary slackness} \end{aligned}$$

Constrained vs multi-objective

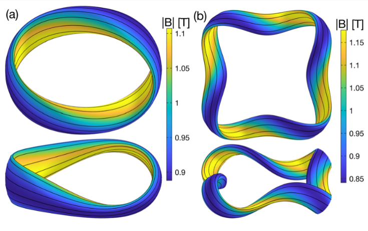
- First-order conditions are *almost* the same
- Can mix and match (constrained multi-objective)
- Multi-objective involves many solves to explore space
- Curse of dimensionality: exploration cost scales exponentially with m

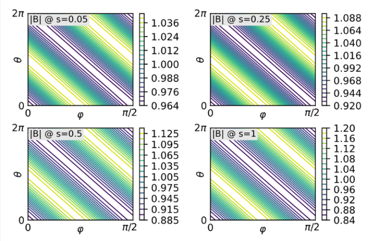
Scalarizing

Find Pareto points via a single-objective optimization problem:

- Linear: $\phi(x) = w^T f(x)$
  - Need to consider stationary points to get full frontier.
  - Uniform weight sampling $\neq$ uniform frontier sampling.
- Projection: $\phi(x) = \sum_i w_i (f_i(x) - f_i^*)^2$
  - Effectively what is done now.
  - Similar tradeoffs to linear scalarization.
- Chebyshev: $\phi(x) = \max_i w_i f_i(x)$
  - Nonsmooth where max is non-unique.
  - Uniform weight $\neq$ uniform frontier sampling.
- ϵ-constraint: $\phi(x) = f_i(x)$, $f_j(x) \leq \epsilon_j$ for $j \neq i$
  - Subproblem is constrained.
  - Can get uniform sampling in components other than i

Example: Quasi-symmetry

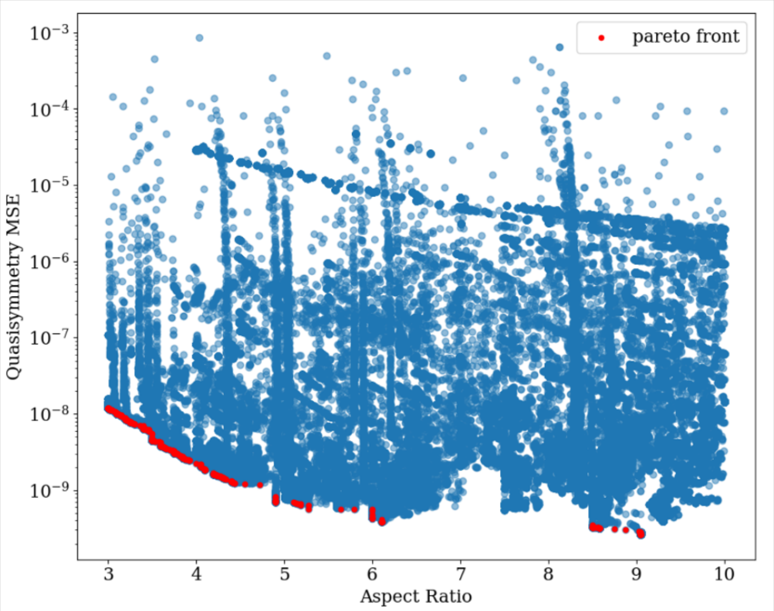




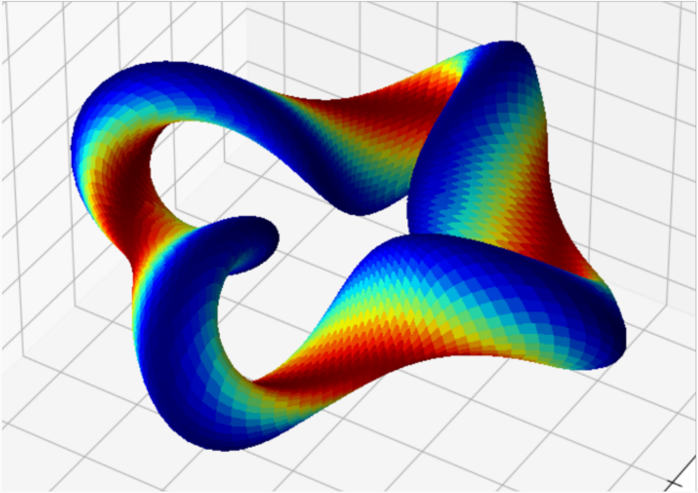
Landreman-Paul QA and QH configurations, optimized with target aspect ratio 6 and 8.

Q: tradeoff between quasisymmetry and aspect ratio? (Padidar, Landreman, Bindel)

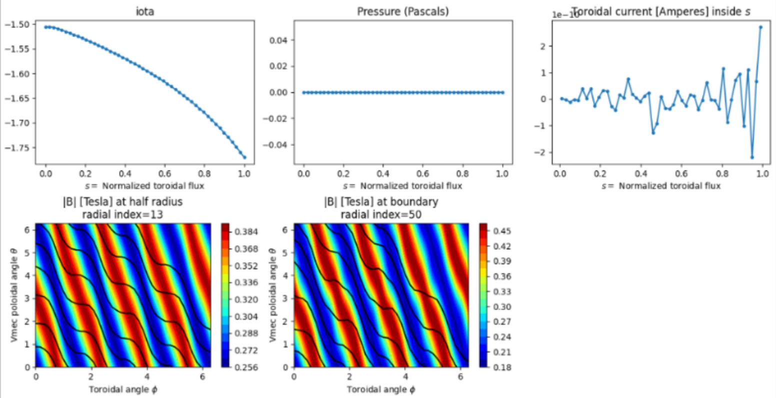
Pareto frontier (QH with 4 field periods)



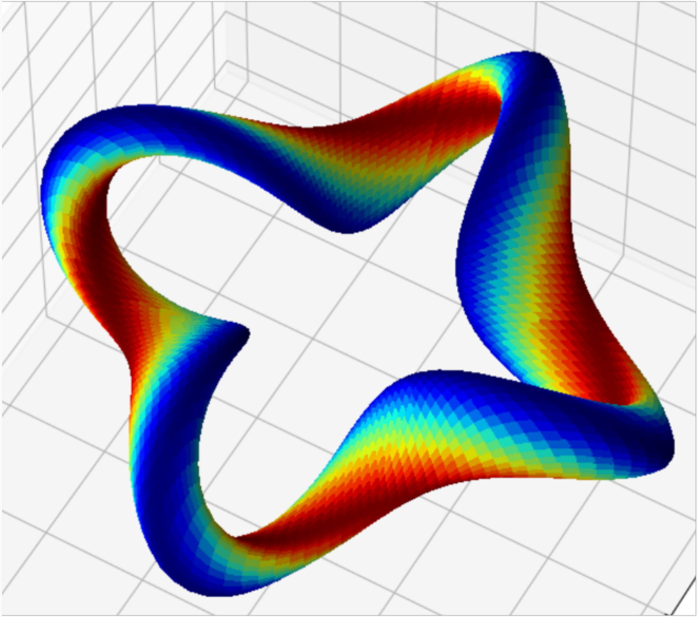
Aspect ratio 3.3



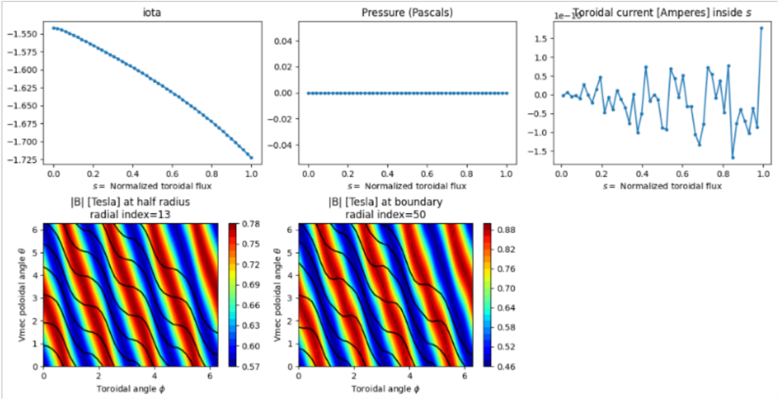
Aspect ratio 3.3



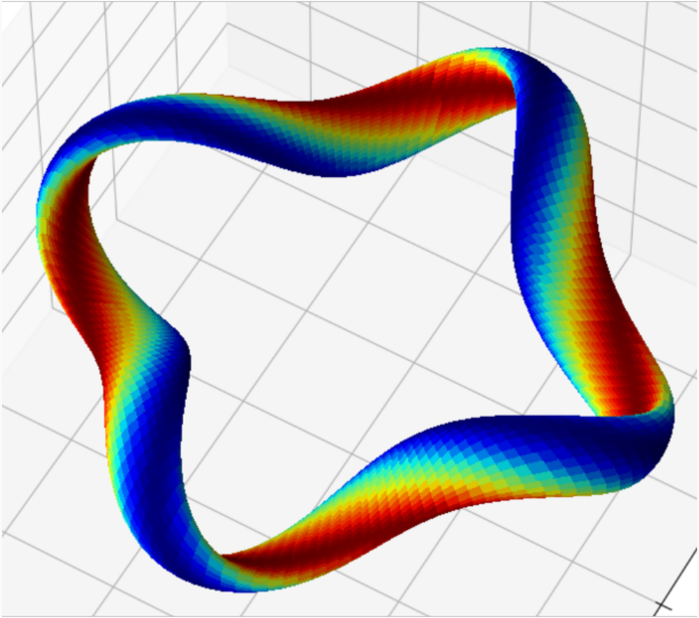
Aspect ratio 5



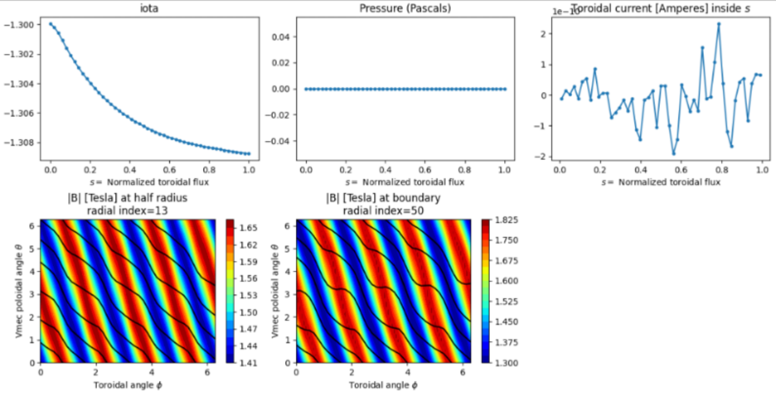
Aspect ratio 5



Aspect ratio 8.67



Aspect ratio 8.67



Continuation

Algorithm in this case: continuation in A

- Start at one Pareto point $(A(x), Q(x))$
- Write stationarity conditions via

$$\begin{aligned} \nabla Q(x) + \lambda \nabla A(x) &= 0 \\ \lambda(A(x) - A^*) &= 0 \\ A(x) &\leq A^* \end{aligned}$$

- Differentiate vs A^* to get tangent direction

$$\begin{bmatrix} \nabla^2 Q(x) + \lambda \nabla^2 A(x) & \nabla A(x) \\ \nabla A(x)^T & 0 \end{bmatrix} \begin{bmatrix} x' \\ \lambda' \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

- Predictor moves a little in tangent direction
- Correct prediction via local solver (e.g. Newton)
- Can re-use Hessians, etc for more efficiency

Which parameterization?

What if Pareto frontier goes vertical?

- Can switch to using Q as continuation parameter
- *Or* use a **pseudo-arclength** parameter
- Generalizations to more than two functions are available (e.g. normal boundary intersection)

Things to ask over coffee

- How many derivatives do I really need?
- Stability objectives or constraint (c.f. Max Ruth on Monday)
- Continuation and numerical bifurcation analysis?
- Other problems where you'd like to understand tradeoffs?