Surrogate Methods for Optimizing Fusion Device Designs

David Bindel

12 December 2019

Department of Computer Science
Cornell University

Acknowledgements

Simons Collaboration: “Hidden Symmetries and Fusion Energy”

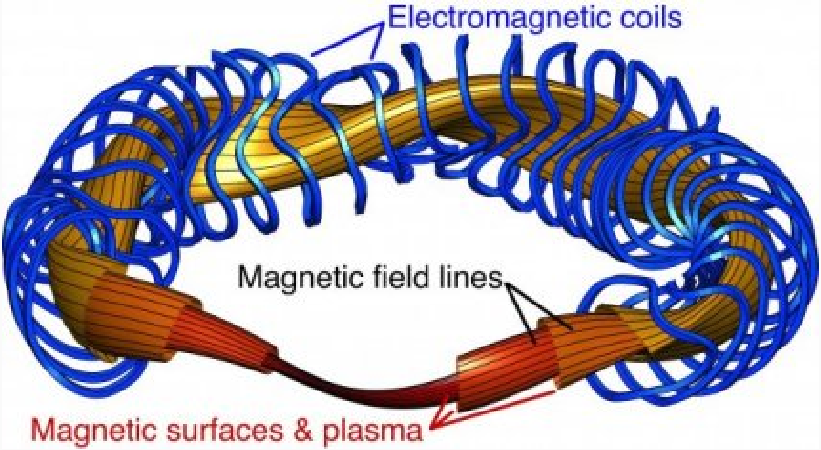
https://hiddensymmetries.princeton.edu/

Cornell group:

- Silke Glas (Simons postdoc)
- Misha Padidar (CAM PhD student)
- Curran Muhlberger (CS lecturer)
- Ariel Kellison (CS MS student)
- Nick Parrilla (predoc student)

with several others on the kernels + optimization side (Eric Lee, Xinran Zhu, Darian Nwankwo, Cameron Ibrahim, Leo Huang, Yuanxi Shen) + former students (David Eriksson, Kun Dong) + collaborators (Peter Frazier, Andrew Wilson) + their students...

Stellarator Concept



What Makes a Good Stellarator?

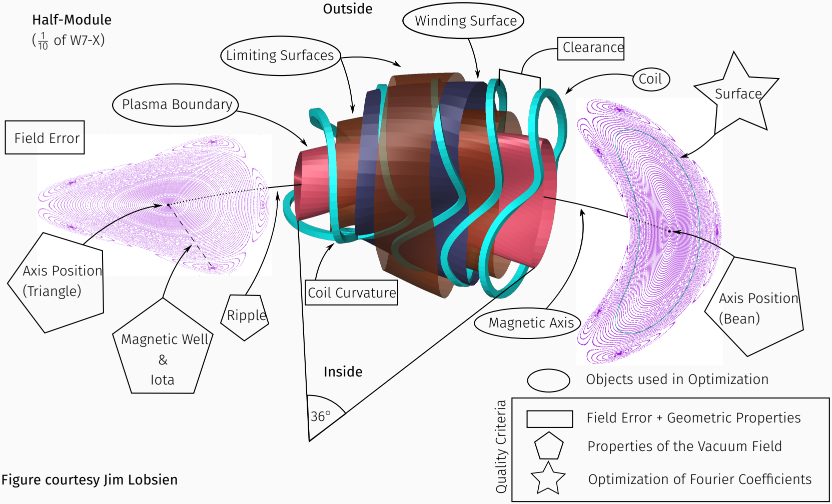


Figure courtesy Jim Lobsien

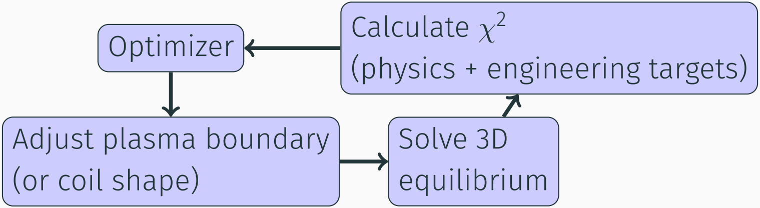
Stellarator Quality Measures

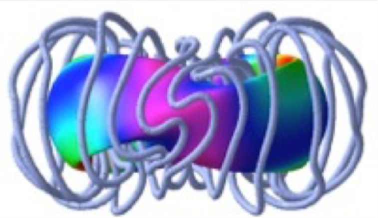
What makes an "optimal" stellarator?

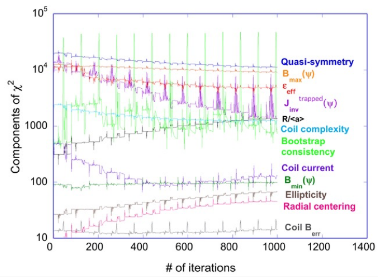
- Approximates field symmetries (which measures?)
- Satisfies macroscopic and local stability
- Divertor fields for particle and heat exhaust
- Minimizes collisional and energetic particle transport
- Minimizes turbulent transport
- Satisfies basic engineering constraints (cost, size, etc)

Each objective involves different approximations, uncertainties, and computational costs.

How Do We Optimize? (STELLOPT Approach)







$$r(\phi,\theta) + iz(\phi,\theta) = \sum \alpha_{m,n} e^{i(m\phi - n\theta)}$$

Challenges

Stellopt: Minimize χ^2 objective over parameter space C:

$$\chi^2(x) = \sum_{k=1}^{m} \frac{J_k(x)}{\sigma_k^2}, \quad J_k(x) = (F_k(x) - F_k^*)^2$$

Solve via Levenberg-Marquardt, GA, differential evolution (avoids gradient information apart from finite differences)

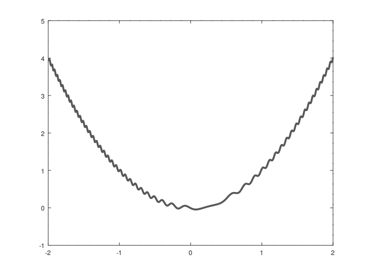
Why doesn't this framework suffice?

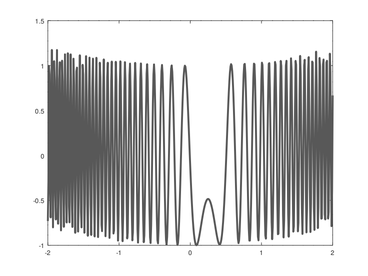
1. Costly and "black box" physics computations
2. Managing tradeoffs (scalarization misses things)
3. Dealing with uncertainties
4. Global search

Tool: Surrogates (*aka* response surfaces, metamodels)

From local to global

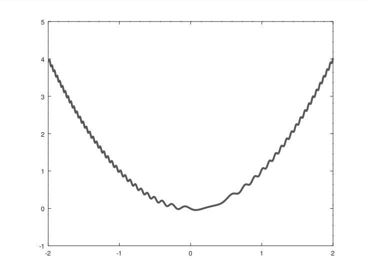
Can *exploit* model for local min; global requires we *explore*.

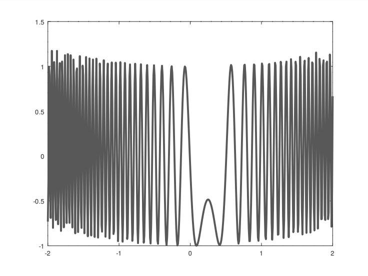




Torn and Zilinskas (1987): For *general* continuous f, convergence to *global* minimum $\implies$ dense sampling

From local to global



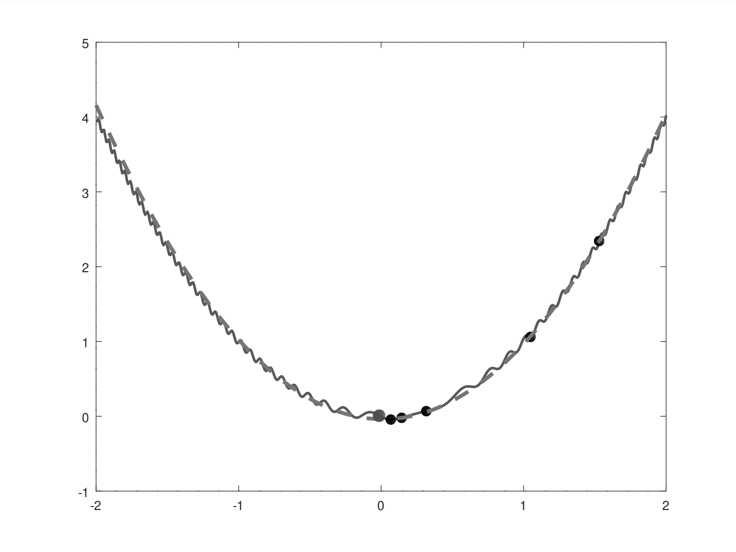


Both problems have local minima that cause issues:

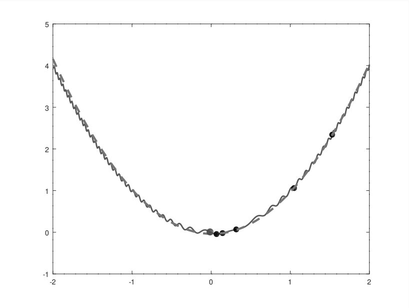
- Left: Good smooth approximations available
- Right: “Glassy” case — little obvious global guidance

These are not equally difficult (especially in high dimensions). Some evidence that we may be (partly) nice.

From local to global: response surfaces



Response surface



Example approach (two-stage):

- Measure $f(x)$ on a sample set (experimental design)
- Fit a *surrogate* or *response surface* by least squares
- Minimize approximating function

Quality depends on model complexity and noise level

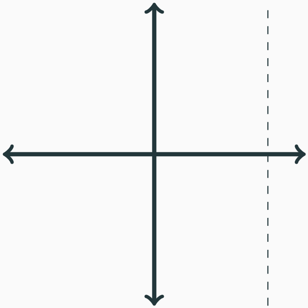
Variations on a theme

Basic idea: Replace expensive f by cheaper $\hat{f}$ (using data)

- Type of surrogate
  - Non-interpolatory (e.g. poly regression, smoothing splines)
  - Interpolatory (e.g. kernel interpolation approaches)
- Surrogate and hyperparameter selection
  - Noise parameters, length scales, etc
- Adaptivity of surrogate
  - Two-stage: Mostly fix surrogate after initial fit
  - One-stage: Continuously update surrogate
- Balancing exploration and exploitation
  - Bayesian interpretations (many types)
  - Frequentist / approximation theoretic
  - Candidate point framework

Focus today: a few things involving kernel-based surrogates.

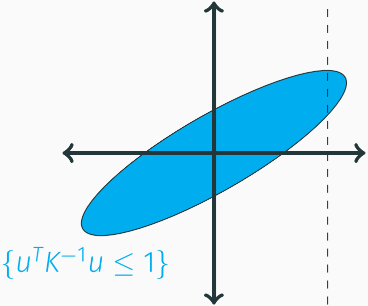
Simple and Impossible



Let $u = (u_1, u_2)$. Given u_1, what is u_2?

We need an assumption! Two different standard takes.

Being Bounded



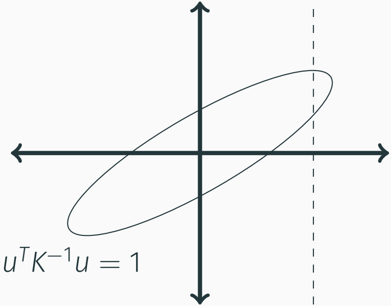
Let $u = (u_1, u_2)$ s.t. $\|u\|^2_{K^{-1}} \leq 1$. Given u_1, what is u_2?

Optimal recovery: $\|u_2 - w\|^2_{S^{-1}} \leq 1 - \|u_1\|^2_{(K_{11})^{-1}}$

$$w = K_{21} K_{11}^{-1} u_1$$

$$S = K_{22} - K_{21} K_{11}^{-1} K_{12}$$

Being Bayesian



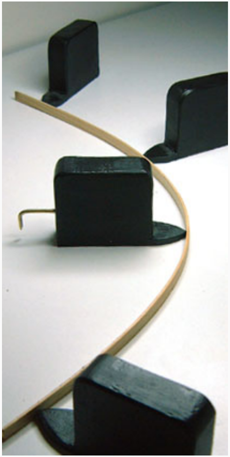
Let $U = (U_1, U_2) \sim N(0, K)$. Given $U_1 = u_1$, what is U_2?

Posterior distribution: $(U_2 | U_1 = u_1) \sim N(w, S)$ where

$$w = K_{21} K_{11}^{-1} u_1$$

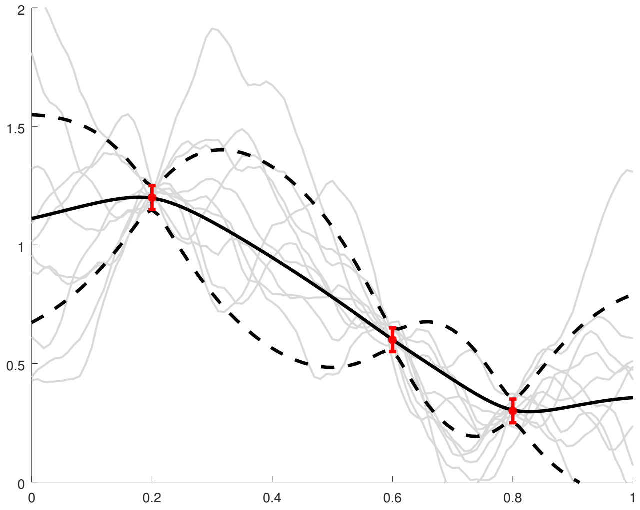
$$S = K_{22} - K_{21} K_{11}^{-1} K_{12}$$

Optimal Recovery: From Energy to Error



http://www.duckworksmagazine.com/03/r/articles/splineducks/splineDucks.htm

Gaussian Processes: Predict with Posteriors



Kernel Interpolation Mechanics

Consider positive definite kernel $k(x, y) = \phi(\|x - y\|)$.
Approximate $s(x) \approx f(x)$ from $f(x_j) = y_j$ via

$$s(x) = \sum_{j=1}^{n} k(x, x_j) c_j$$

Interpolation conditions

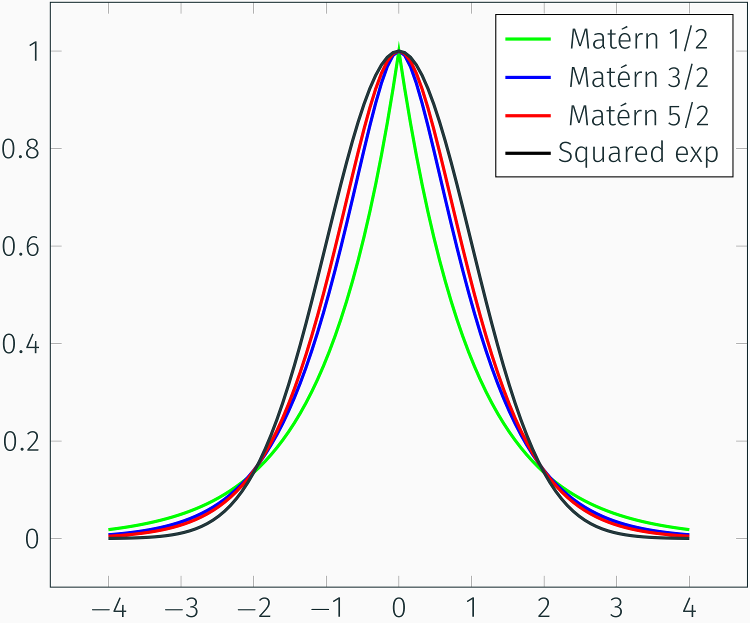
$$K_{XX} c = y$$

where $[K_{XX}]_{ij} = k(x_i, x_j)$. Error at x associated with

$$v(x) = k(x, x) - k_{xX} K_{XX}^{-1} k_{Xx}$$
$$k_{xX} := k_{Xx}^T = \begin{bmatrix} k(x, x_1) & \dots & k(x, x_n) \end{bmatrix}$$

Call $v(x)$ the *predictive variance at x* (GP setting) or *squared power function at x* (optimal recovery setting).

Common Kernels



Optimal Recovery vs Bayesian Approach

Different error frameworks:

- Optimal recovery / variational approach

$$|f(x) - s(x)| \leq \sqrt{v(x)\left(|f|_{\mathcal{H}}^2 - |s|_{\mathcal{H}}^2\right)}$$

  Minimize worst-case error subject to regularity condition (bound on norm of f in RKHS).
- Bayesian approach

$$(f(x)\,|\,f(x_i) = y_i) \sim N(s(x), v(x))$$

  Uncertainty / error characterized via Gaussian
- Approaches coincide when data/predictions are *linear*
- Different error philosophies $\implies$ different regularization, kernel selection, ...
- Really different for nonlinear data/predictions

Details Under the Rug

Several topics for a different time:

- Kernel selection and hyper-parameter tuning
- Mean fields and combinining GPs with fixed bases
- Multi-output GPs and prediction with co-variates
- Scaling to large numbers of data points
- Dimensionality issues

What matters for now: prediction + uncertainty

Bayesian Optimization

- Assume prior $f \sim \mathrm{GP}(\mu, k)$
- Condition on observations so far:

$$
\begin{aligned}
(f \,|\, f(x_i) = y_i) &\sim \mathrm{GP}(\hat{\mu}, \hat{k}) \\
\hat{\mu}(x) &= \mu(x) + k_{xX} c \\
K_{XX} c &= y - \mu \\
\hat{k}(x, y) &= k(x, y) - k_{xX} K_{XX}^{-1} k_{Xy}
\end{aligned}
$$

- Choose next point to optimize an *acquisition function*, e.g.

$$
\begin{aligned}
\mathrm{EI}(x) &= \mathbb{E}[f_{\mathrm{best}} - f(x)]_+ \\
&= (f_{\mathrm{best}} - \hat{\mu}(x))\Phi(Z) + \hat{\sigma}(x)\phi(Z), \\
Z &= \frac{f_{\mathrm{best}} - \hat{\mu}(x)}{\hat{\sigma}(x)}
\end{aligned}
$$

- Many other acquisition functions (some non-Bayesian)

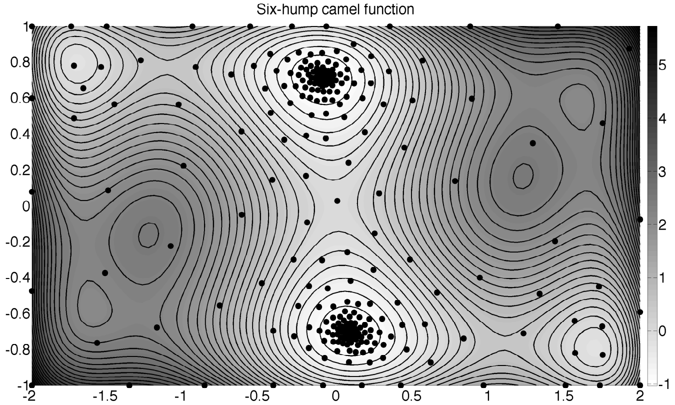
Variational Approach

- Assume $|f| \leq C$
- Obtain lower bounds based on observations so far:

$$\begin{aligned} f(x) &\geq s(x) - \sqrt{v(x)\,(C^2 - |s|^2)} \\ s(x) &= k_{xX}c = \sum_i k(x, x_i)c_i \\ v(x) &= k(x,x) - k_{xX}K_{XX}^{-1}k_{Xx} \\ c &= K_{XX}^{-1}y \\ |s|^2 &= c^T y \end{aligned}$$

- Choose next point to minimize lower bound
- Provably globally convergent (Eriksson and B)
- V similar to UCB/LCB acquisition in Bayesian case

Exploration vs Exploitation (Eriksson and B)



Linearity of Gaussians

Linear functionals of a GP are Gaussian:

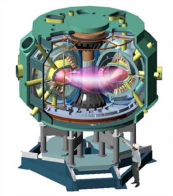
- Use to incorporate derivatives in BO
- Use to incorporate integral constraints
- What else can we do with this?

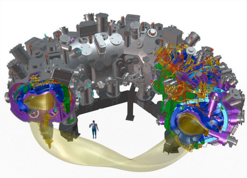
Optimization Under Uncertainty

Low construction tolerances:

- NCSX: 0.08%
- Wendelstein 7-X: 0.1% – 0.17%

Want: higher tolerances as coil optimization goal!

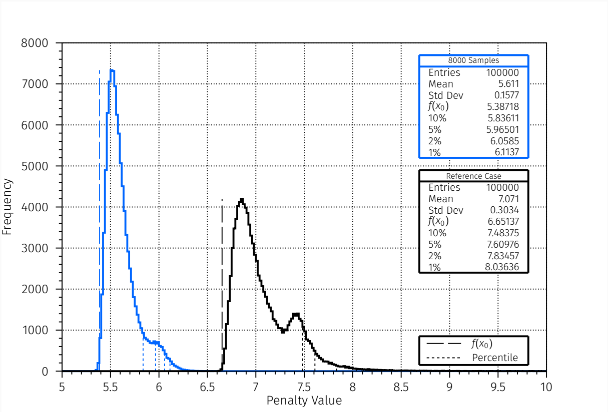




Also want tolerance to

- Changes to control parameters
- Uncertainty in physics or model

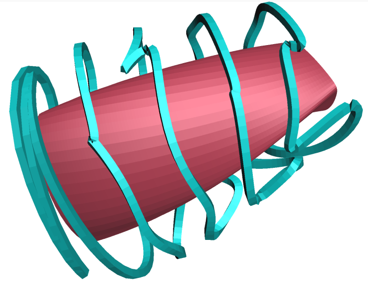
Monte Carlo Approach



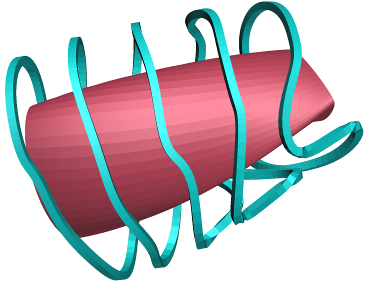
Robustness & average performance significantly improved

[Lobsien, Drevlak, Sunn Pedersen]

Example



Classic Coil Optimization (Deviation 0 mm, Penalty = 4.19)



Stochastic Optimization (Deviation 2.5 mm, Penalty = 2.24)

Risk-Neutral OUU: Easy Case

- Expected objective $\hat{J}(x) = \mathbb{E}_U[J(x+U)]$, $U \sim \omega$
- Model J from GP prior with mean 0, kernel $k(x,y) = \phi(\|x-y\|)$
- Posterior for J has mean and kernel

$$\mu(x) = \sum_i c_i \phi(\|x - x_i\|) \qquad \hat{k}(x,y) = \phi(0) - k_{xX} K_{XX}^{-1} k_{Xy}$$

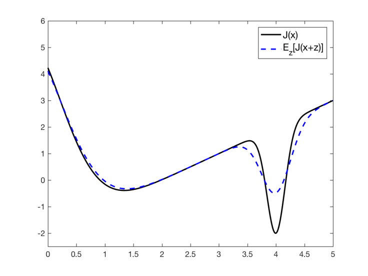
- Can derive associated posterior for $\hat{J}$:

$$\check{\mu}(x) = \sum_i c_i \tilde{\phi}(\|x - x_i\|) \qquad \check{k}(x,y) = C - \tilde{k}_{xX} K_{XX}^{-1} \tilde{k}_{Xy}$$

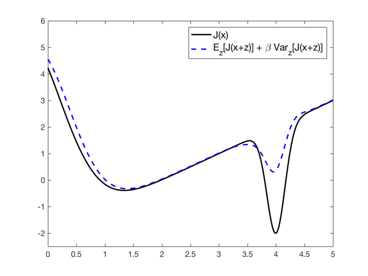
$$\tilde{\phi} = \phi * \omega, \qquad C = \int \mathcal{F}(\phi)\mathcal{F}(\omega)^2$$

- Now apply usual BO framework
- More general uncertainty: apply BQO (Frazier and Toscano)

Risk Neutral vs Risk Averse



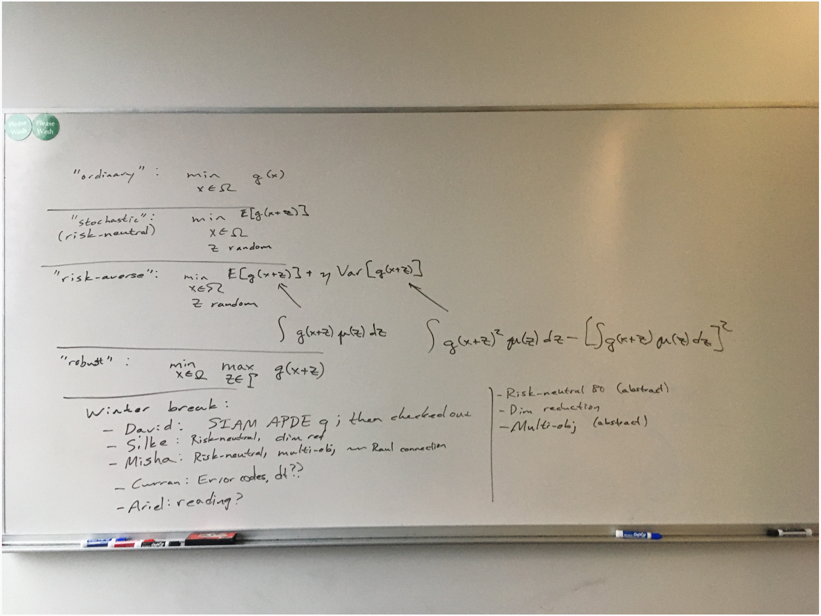
Risk-Neutral Objective



Risk-Averse Objective

- Risk averse objective is *not* linear in *J* (not a GP model)
- Can still do BO (initial work with Peter Frazier)

Where We Are



"ordinary" : min x∈Ω g(x)
"stochastic": (risk-neutral) min x∈Ω z random E[g(x+z)]
"risk-averse": min x∈Ω z random E[g(x+z)] + η Var[g(x+z)]
∫ g(x+z) μ(z) dz
∫ g(x+z)^2 μ(z) dz − [∫ g(x+z) μ(z) dz]^2
"robust" : min x∈Ω max z∈Γ g(x+z)
Winter break:
– David: SIAM APDE g; then checked out
– Silke: Risk-neutral, dim red
– Misha: Risk-neutral, multi-obj, ~ Raul connection
– Curran: Error codes, dt??
– Ariel: reading?
– Risk-neutral 80 (abstract)
– Dim reduction
– Multi-obj (abstract)

Next Steps

- We can call Stellopt as a black box (via RPC)
- We have initial tests with a low-dimensional problem
- Know how to do some OUU
  - Need to code the thing up
  - Need to talk to Jim-Felix Lobsien for a problem!
  - Risk averse form is a bit further behind
- Interested in multi-objective as well
  - Need to integrate some standard methods
  - Modeling user implicit preferences? (Raul Marban and Peter Frazier)
- Looking at kernel surrogates beyond standard GP/RBF
  - Incorporate non-Gaussian effects, upper/lower bounds, etc
  - Allows some error analysis with less regularity

Ideally: also looking for applications from you!