Deterministic Langevin Optimization

SIAM OP23

Jamie Sullivan
(work w/ Uroš Seljak)
UC Berkeley

Berkeley
UNIVERSITY OF CALIFORNIA

DOE
CSGF

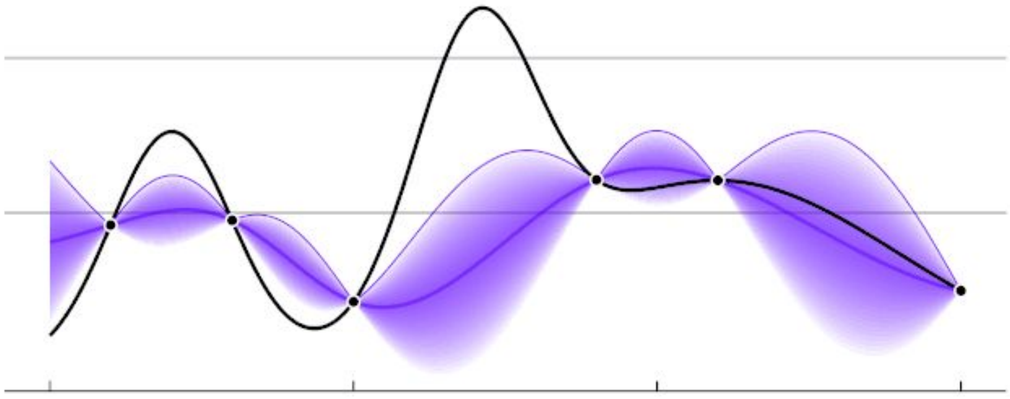
BERKELEY CENTER *for*
COSMOLOGICAL PHYSICS

Active Learning

Strategy for expensive functions:

1. Think very carefully about choosing a domain point
2. Evaluate the function at the top candidate point
3. Build upon success or learn from mistakes

Iterate!

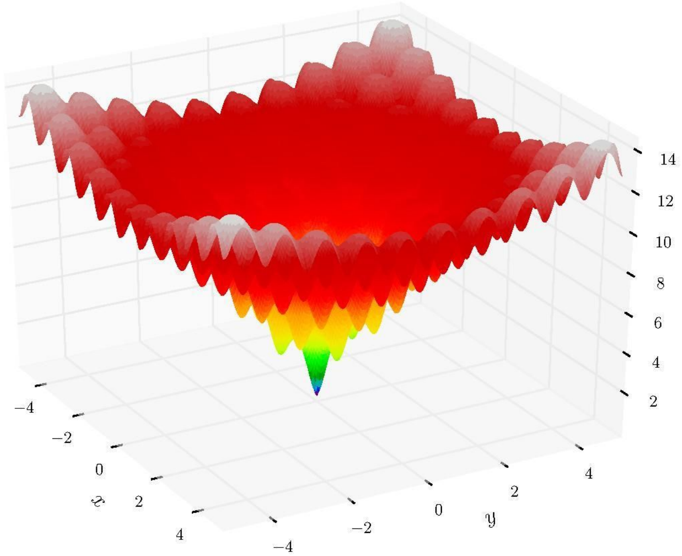


Bayesian Optimization

Bayesian optimization (BO) is a strategy for global optimization of expensive black-box functions

Local methods can fail on rugged or multi-modal objectives

Spaces of moderate dimension ($d<100$)

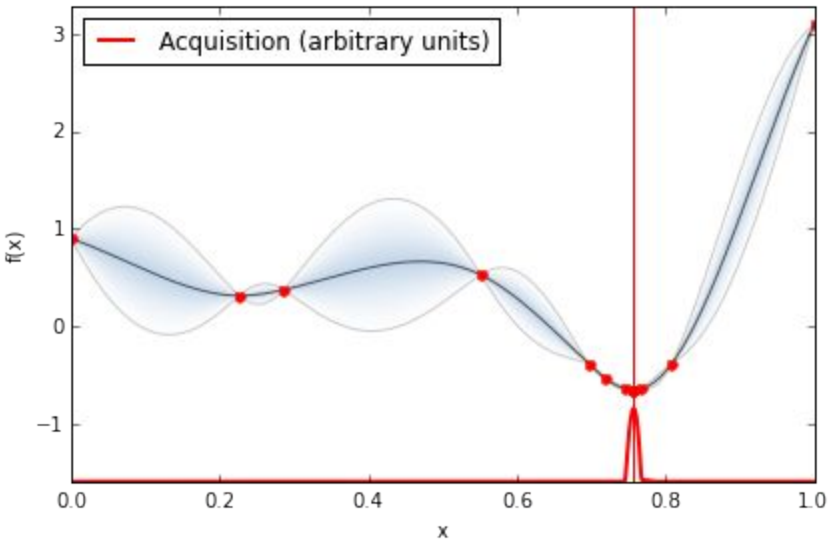


Bayesian Optimization

1. Surrogate model: *s(θ)* - approximates the function
2. Acquisition function (AF): weights exploration vs exploitation to select future points

Gaussian Processes (GPs) are non-parametric interpolators with uncertainty attached

Typically *s(θ)* is a Gaussian Process mean, and the *AF* uses GP uncertainty



DLO - Acquisition Function

GP comes with a "natural" uncertainty for the AF

We instead use a density estimate for uncertainty inspired by the deterministic Langevin equation:

$$\theta_{t+1} = \theta_t + v\epsilon = \theta_t + \frac{d}{d\theta}[\beta f(\theta_t) + V_t(\theta_t)(t)]\epsilon$$

DLO - Acquisition Function

GP comes with a “natural” uncertainty for the AF

We instead use a density estimate for uncertainty inspired by the deterministic Langevin equation:

stochastic update

$$\theta_{t+1} = \theta_t + v\epsilon = \theta_t + \frac{d}{d\theta}[\beta f(\theta_t) + V_t(\theta_t)(t)]\epsilon$$

parameter vector
“particle motion”

Write velocity as the gradient of the combination of log target and density

DLO - Acquisition Function

GP comes with a “natural” uncertainty for the AF

We instead use a density estimate for uncertainty inspired by the deterministic Langevin equation:

$$\theta_{t+1} = \theta_t + v\epsilon = \theta_t + \frac{d}{d\theta}[\beta f(\theta_t) + V_t(\theta_t)(t)]\epsilon$$

$$\theta_{t+1} = \arg\max_\theta[\beta f(\theta) + V_t(\theta)] = \arg\max_\theta \ln \frac{\exp(\beta f(\theta))}{q_t(\theta)}$$

Formulate as optimization problem!

$$q(\theta) \equiv e^{-V(\theta)}$$

DLO - Acquisition Function

GP comes with a “natural” uncertainty for the AF

We instead use a density estimate for uncertainty inspired by the deterministic Langevin equation:

$$\theta_{t+1} = \theta_t + v\epsilon = \theta_t + \frac{d}{d\theta}[\beta f(\theta_t) + V_t(\theta_t)(t)]\epsilon$$

$$\theta_{t+1} = \arg\max_{\theta}[\beta f(\theta) + V_t(\theta)] = \arg\max_{\theta} \ln \frac{\exp(\beta f(\theta))}{q_t(\theta)}$$

target

density estimate

Normalizing Flows

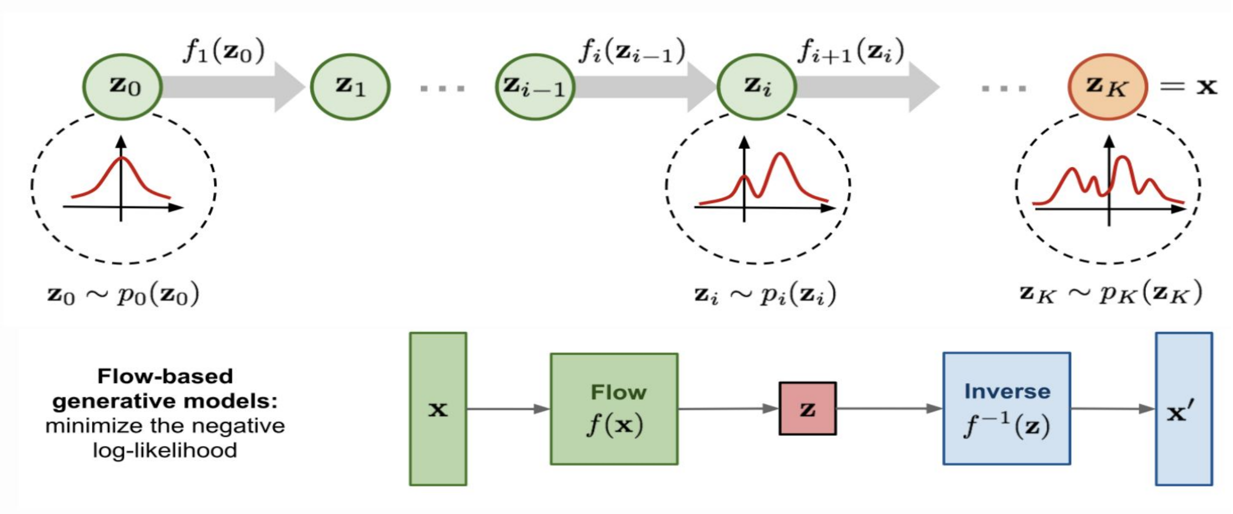
NFs give a bijective map from a base distribution to a target

(Rezende & Mohamed 15, Papamarkios++19)

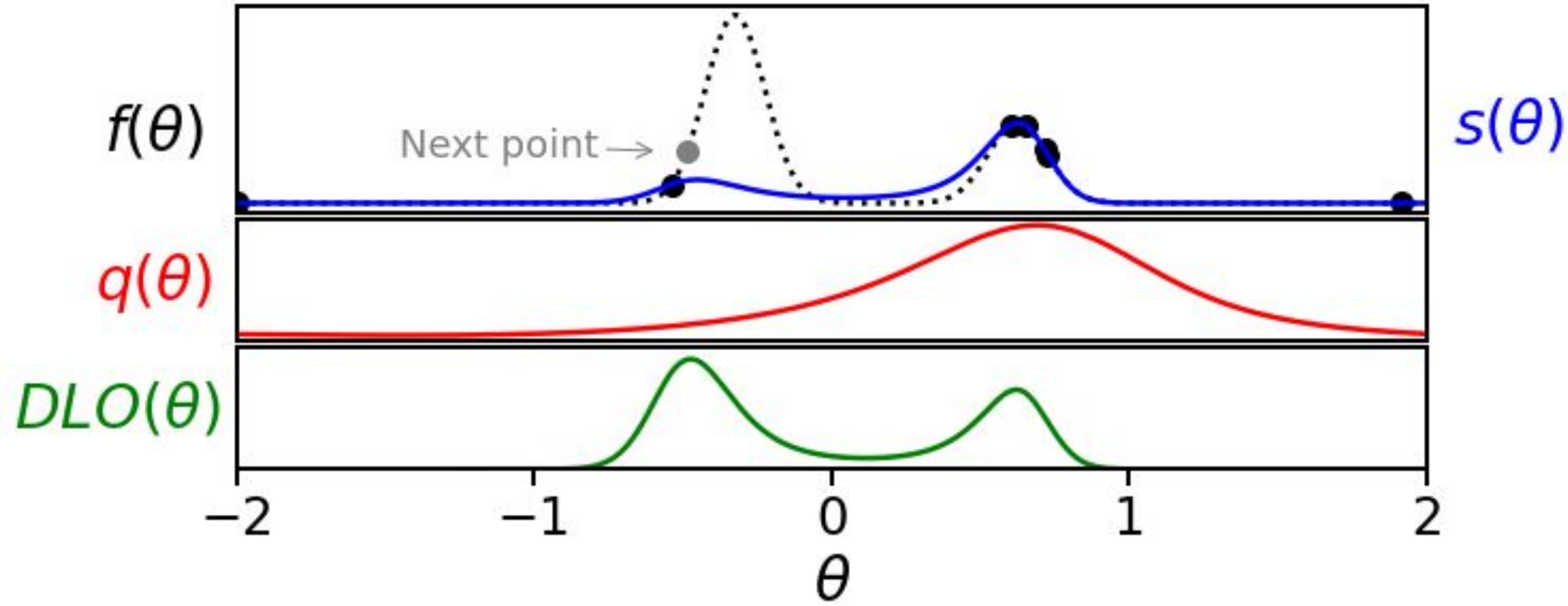
Fast sampling and evaluation of approximate density

We use a Sliced Iterative Normalizing Flow

(Dai & Seljak 20)



DLO - almost Bayesian Optimization



Schematic Algorithm

The flow allows us to search for new points in *latent space*

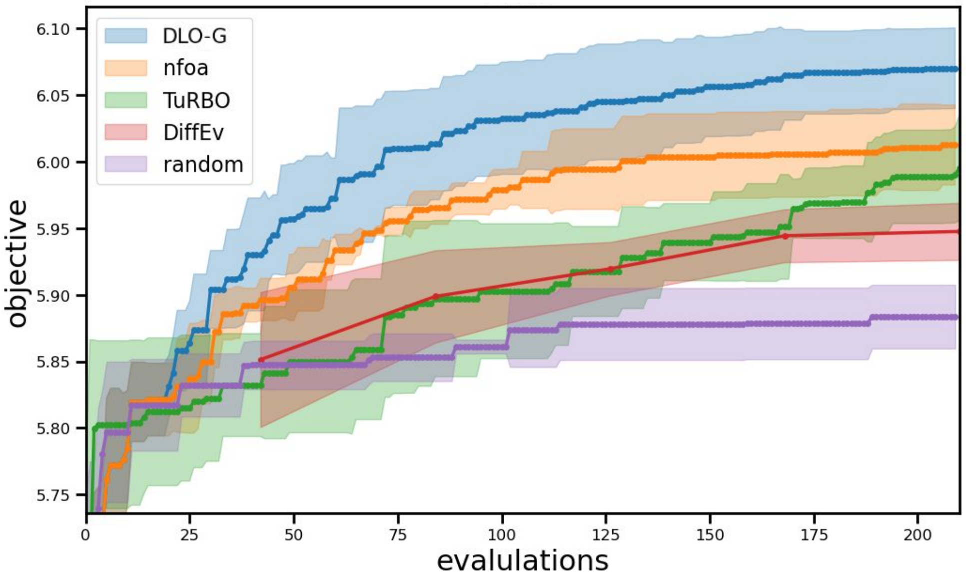
Algorithm 1: Schematic Version

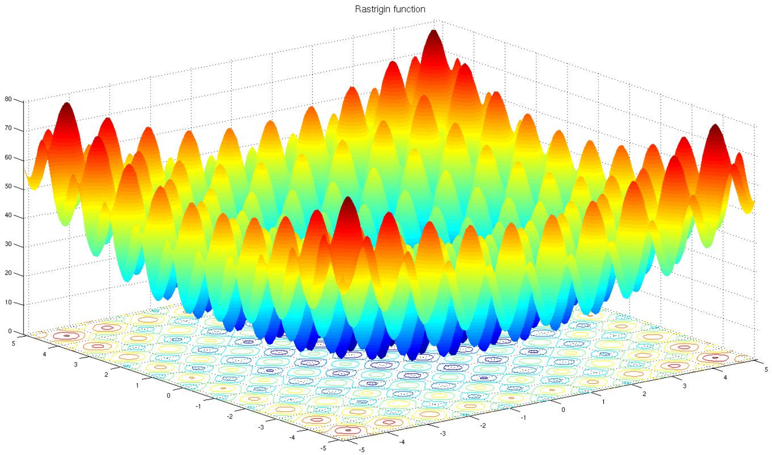
Evaluate $f(\theta_0)$ at initial points.
Assign call budget N, which sets the annealing levels N_β.
for $i < N_\beta$ **do**
- Fit NF q_{uw} to obtain unweighted sample density.
- Fit surrogate q_w to annealed objective values $\beta \log p(\theta)$.
- Locally maximize the acquisition function $AF(\theta)$
- Evaluate $f(\theta_{i+1})$ and update β.

end

DLO Results: Test Functions

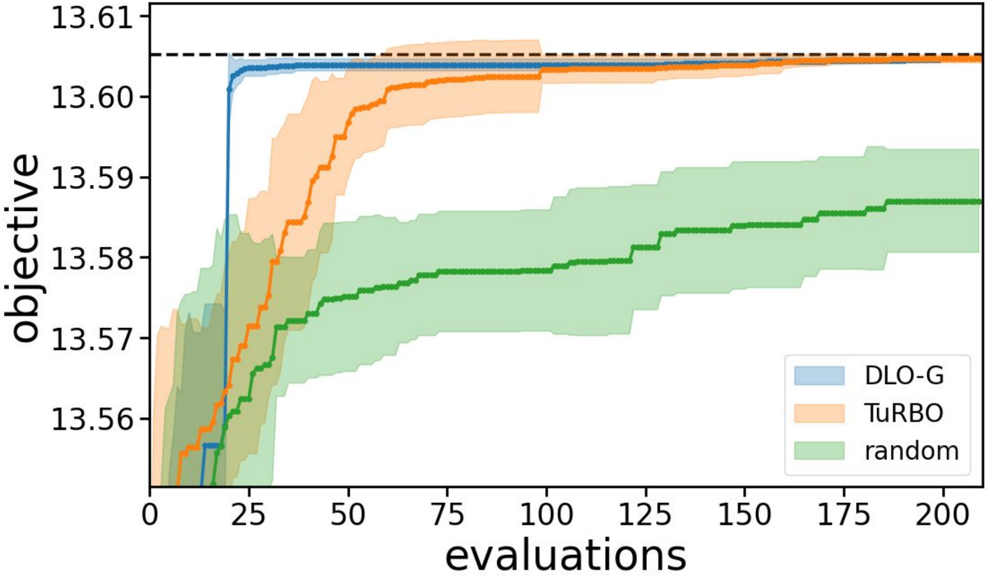
Rastrigin & Rosenbrock objectives in 10d

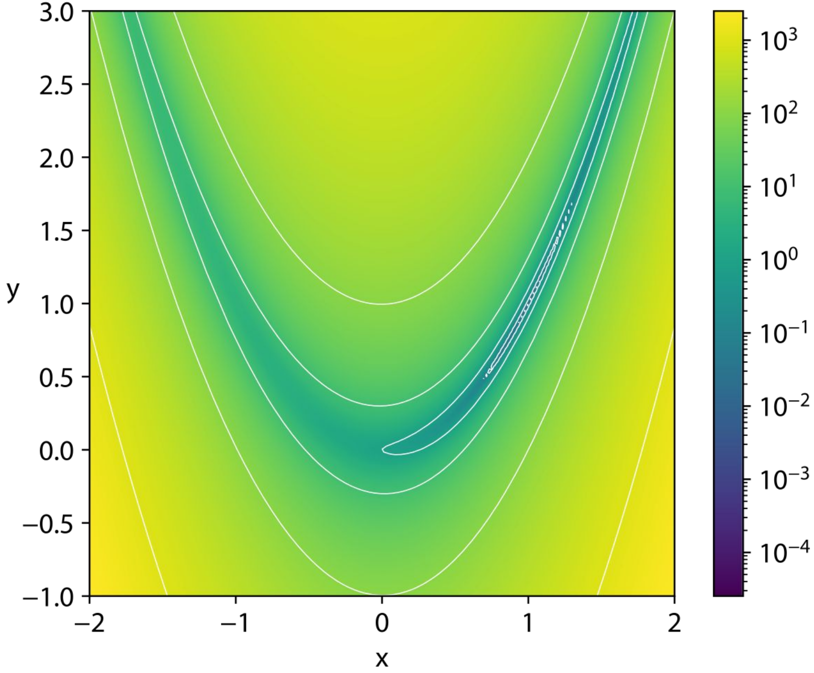




DLO Results: Test Functions

Rastrigin & **Rosenbrock** objectives in 10d

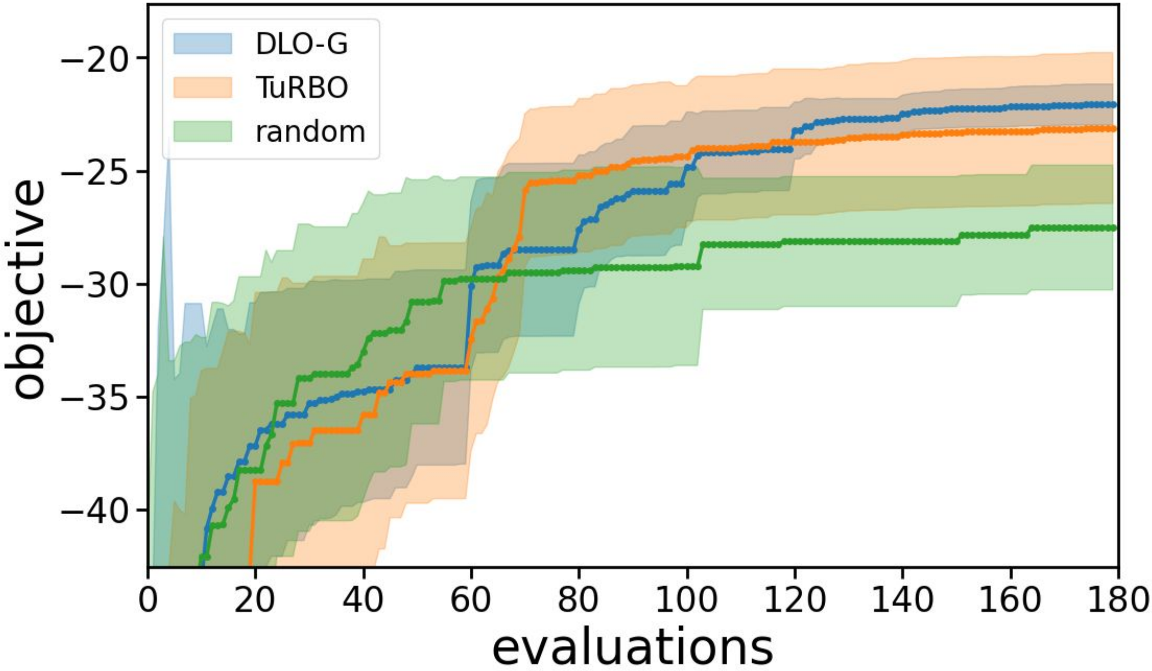




DLO Results: Applied example

Cosmology application:
Luminous Red Galaxy clustering

11-d posterior
Inference problem



Also competitive for
ML hyperparameter optimization

Choice of Surrogate

For lower-*d*, use GP, but DLO works with NNs

Runtime for GP becomes intractable with d

NN instead can save wall-clock time:

d	2	5	10	20	50
Evaluation:					
DLO-GP	0.02	0.03	0.07	0.15	6.07
DLO-NN	0.02	0.02	0.04	0.10	0.53
Fitting:					
DLO-GP	0.17	0.26	0.44	1.71	46.65
DLO-NN	0.03	0.04	0.06	0.20	2.84

DLO Summary

NF density can replace GP uncertainty

Success on moderately high-dimensional targets

Other surrogates scale to higher *d* / larger datasets than GP

Extra Slides

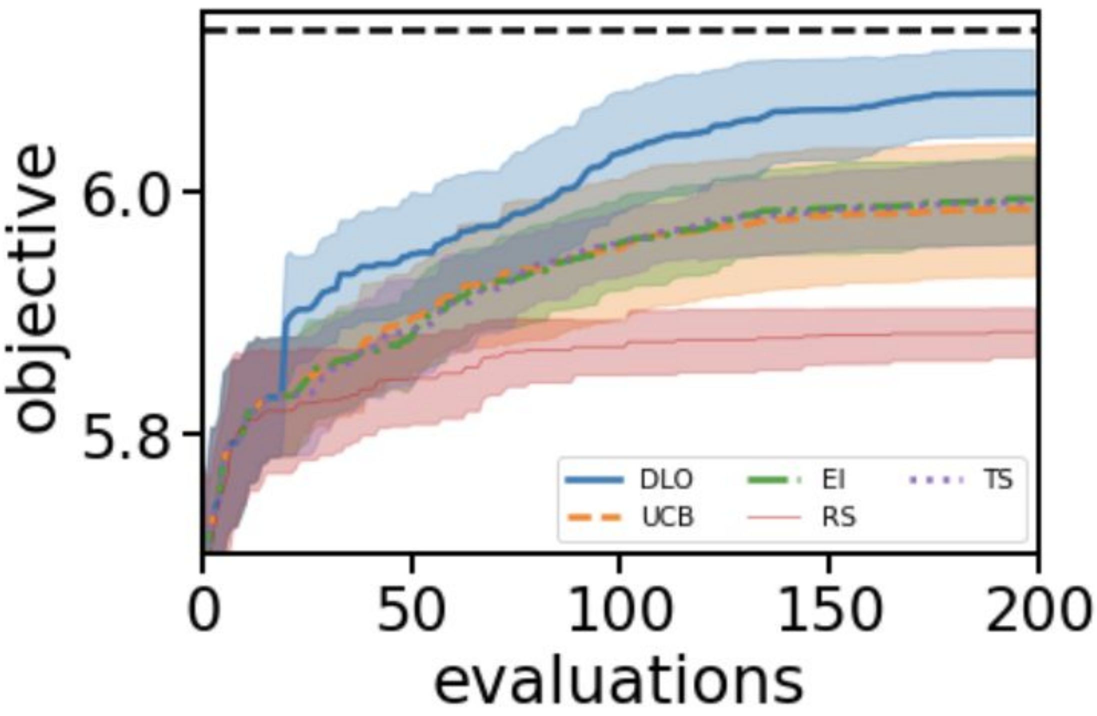
Acquisition Functions

DLO

Upper Confidence Bound

Expected Improvement

Thompson Sampling

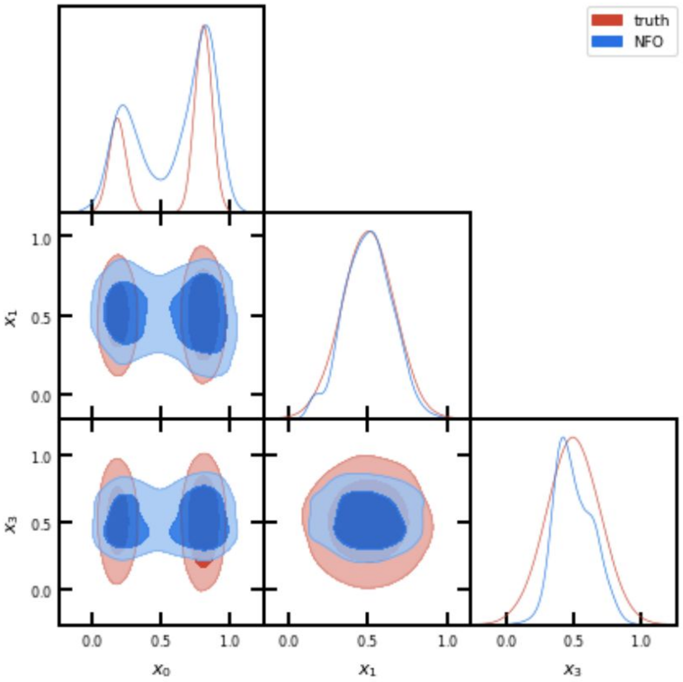


DLO as MCMC burn-in …

DLO provides a good starting point for sampling

100 importance-weighted samples is already qualitatively correct in 10d

Corrected surrogate steps perform even better on harder problems



Full Algorithm

Algorithm 1 Deterministic Langevin Optimization

Evaluate $f(\theta_1), .., f(\theta_{N_I})$ at N_I initial points; select initial annealing level β_0, rescale the input θ domain to $[0, 1]^d$.

Assign a call budget N, fix the hyperparameters N_β, R, dR.

for $i < N_\beta$ **do**

    Estimate the normalizing flow density $q_i(\theta)$ from $\theta_1, .., \theta_t$.

    Fit the surrogate $s_i(\theta, \beta_i))$ from $f(\theta_1), ..f(\theta_t)$ to annealed objective values.

    Create proposal samples in $[0, 1]^d$ *and* in the latent space of q_t drawing from Gaussian spheres of radius R around the highest $\mathrm{DLO}(\theta_j), j = 1...t$.

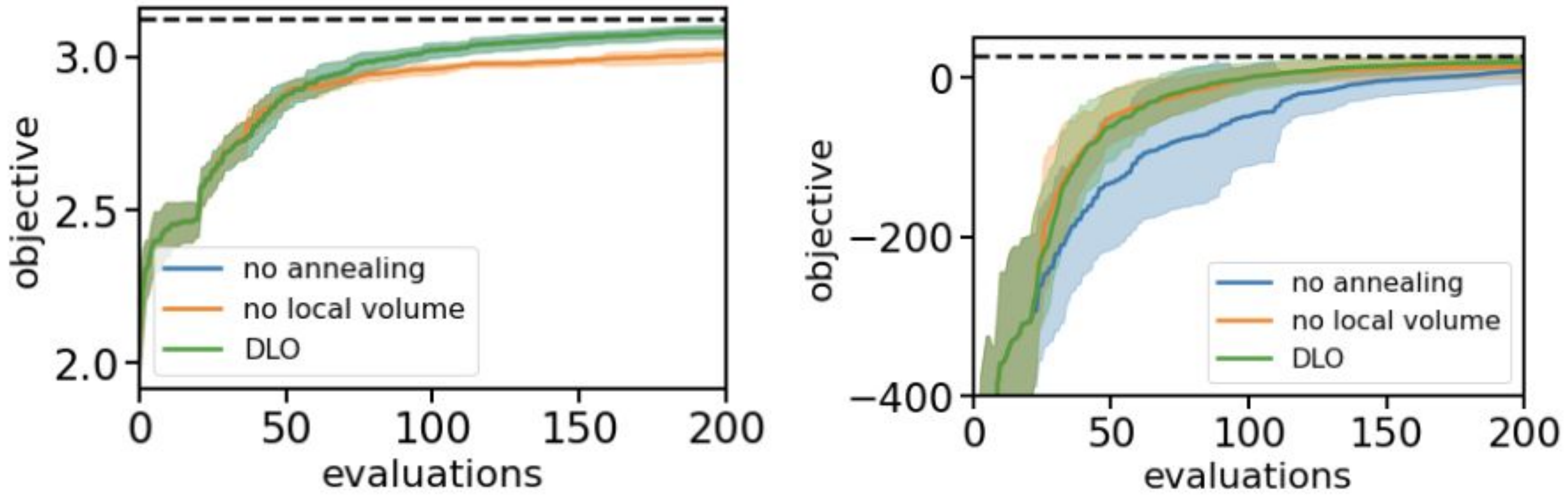
    Locally maximize the acquisition function $\mathrm{DLO}(\theta)$ from N_{sample} proposal draws to obtain the next batch of $\theta_{t+1}, .., \theta_{t+B}$ to evaluate.

    Evaluate $f(\theta_{t+1}), ..f(\theta_{t+B})$ and update β_i.

end for

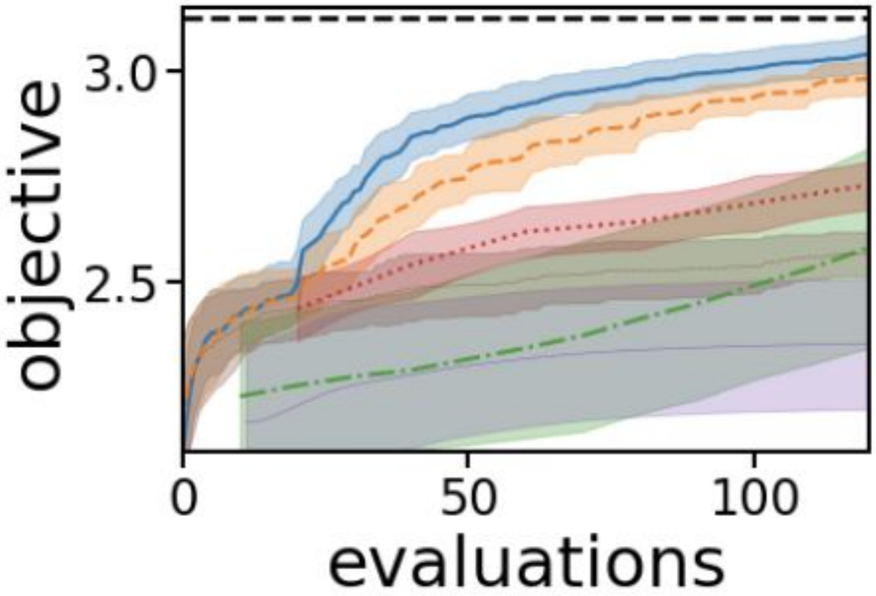
Local Exploration & Annealing Ablation

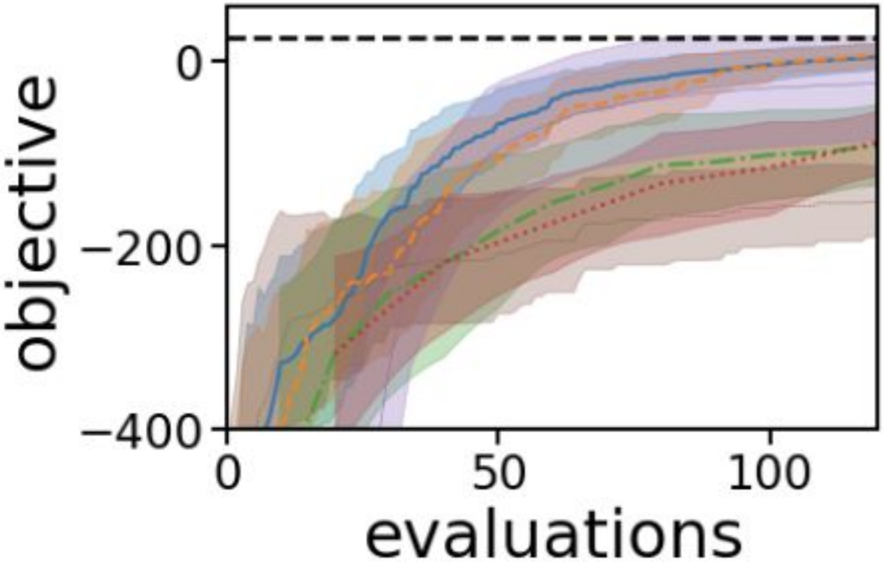
10-*d* Ackley and Correlated Gaussian



More targets: Ackley & CG (d=10)

10-*d* Ackley and Correlated Gaussian





DLO Strategy