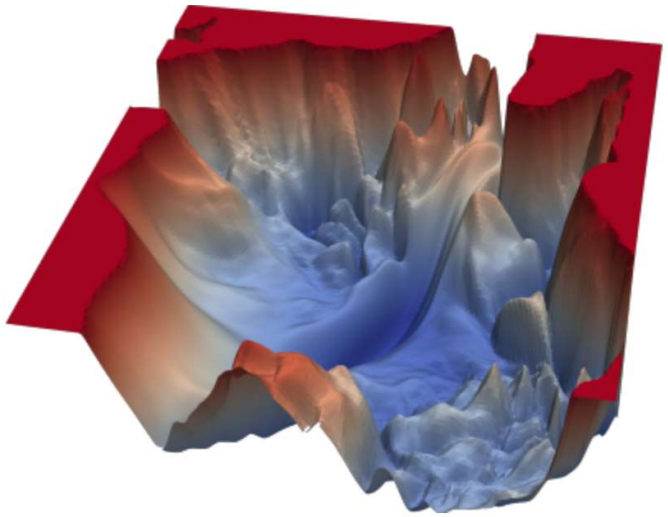
To appear (early 2022)
Offshoot of discussions w/ J. Batson, Y. Kahn, D. Roberts on inflation and optimization

For many problems one needs to minimize an objective (`Loss') function V, descending a generally non-convex high dimensional landscape.

--data analysis/machine learning

-- PDE solving, Loss = $\sum(PDEs)^2 + (boundary\ conditions)^2$: want global min

Gradient descent methods and variants can work well, but sometimes get stuck at a (high) local minimum and/or don't sample all desired solutions.



Early U cosmology: models for descending a potential landscape V.

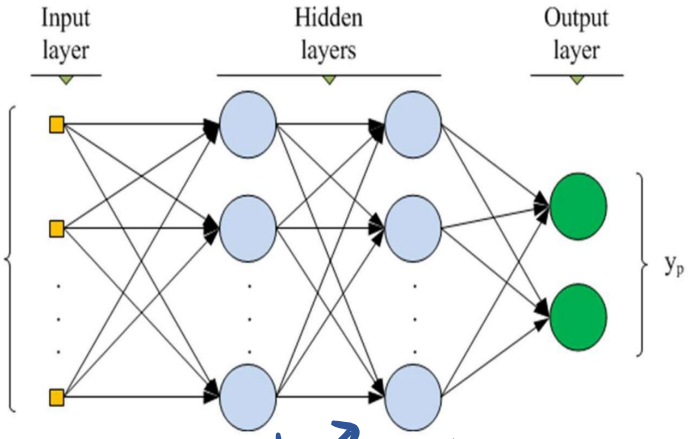
--Example: DBI: relativistic speed limit → 0 as $V \to 0$ *without friction, consistent with energy conservation*

cf Relativistic Gradient Descent Franca et al '19 (with constant speed limit)

Lightning intro to NN's for PDE solving

Cf Lagaris, Likas, Fotiadis '97,..., **Many talks** ML for Calabi–Yau metrics, Ricci flow, etc.; DL S T'21: fields in hyperbolic compactification

Points sampled from domain of PDE



Output functions (ansatzes for functions being solved for)

$\sigma(W \cdot X + b)$

denote as $\vec{\theta}$ (NN parameters)

Repeated application builds up nonlinear output functions/ansatzes

Then form loss functional: e.g.

$$\sum_{pts,eqs} (PDEs)^2 + (boundary\ conditions)^2$$

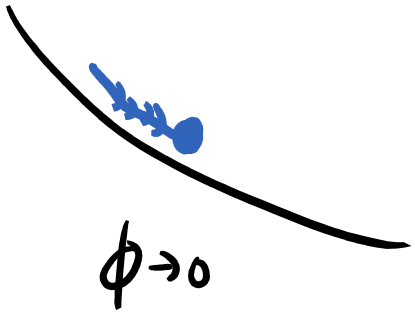
Descend the loss landscape via gradient descent or generalizations

Early U inflation requires nearly constant $V(\phi)$

- Slow roll (flat potential, Hubble friction dominates)
- Interactions slow the field, e.g. DBI inflation: speed limit ϕ-dependent

$$S = -\int d^4x \left\{ \frac{\phi^4}{\lambda}\sqrt{1-\frac{\lambda\dot{\phi}^2}{\phi^4}} + \Delta V(\phi) \right\}$$

Testable (falsifiable(?)) via non-Gaussianity (≃equilateral shape) [`String Data']



$f_{NL}^{DBI} = 14 \pm 38$ Planck

$f_{NL}^{local} = -0.9 \pm 5.1$; $f_{NL}^{equil} = -26 \pm 47$; and $f_{NL}^{ortho} = -38 \pm 24$ (68 % CL, statistical)

Distinct behavior and predictions from slow roll

Non-gravitational version conserves energy (no friction), only stopping at V=0

$$S = -\int V(\vec{\theta})\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{V(\vec{\theta})}} \qquad \pi_i = \frac{\partial L}{\partial \dot{\theta}^i} = \frac{\dot{\theta}_i}{\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{V}}}$$

$$H = \frac{V}{\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{V}}} = \sqrt{V(V + \vec{\pi}^2)} \equiv E = constant$$

$\Rightarrow$ **Cannot** stop at local min, even without stochastic noise (but can get stuck in orbit). **Cannot** overshoot V=0.

Distinct behavior from gradient descent

Phase space volume strongly dominated near global minimum:

$$Vol(\mathcal{M}) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int d^n\theta \int d\tilde{\pi}\tilde{\pi}^{n-1}\delta(\sqrt{V(V + \tilde{\pi}^2)} - E) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int d^n\theta \frac{E}{V}\left(\frac{E^2}{V} - V\right)^{\frac{n-2}{2}}$$

Many variations on this theme, e.g.

- `Log inflation' mechanism with log rather than square root branch cut ←→ speed limit. From integrating out flavor fields:

$$\Gamma_{1PI} = \int a^3 \left\{ \frac{1}{2}\dot{\phi}^2\left(1+\frac{\chi^2}{M_*^2}\right)+\frac{1}{2}(\partial\chi)^2 - V(\phi) - \Delta V_{eff}(\chi) - \frac{N_f}{2}\int_H^{M_{UV}} \frac{d^4k_E}{(2\pi)^4}\log\left(1-\frac{\dot{\phi}^2/M_*^2}{k_E^2 - i\epsilon}\right)\right\}$$

(w/Mathis, Mousatov, Panagopoulos '20):

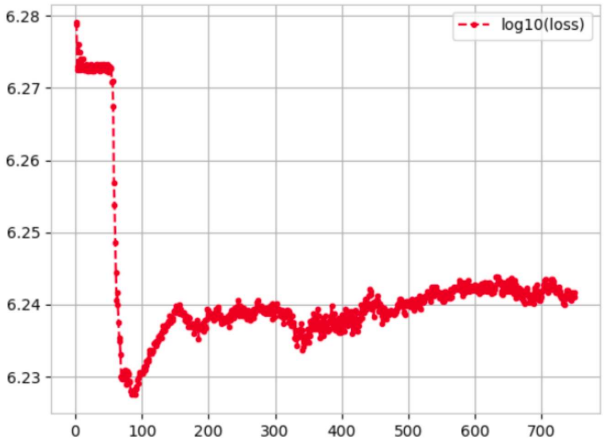
- 2-derivative action with mass ~ 1/Loss

As an energy conserving dynamical system in a rich loss landscape (without symmetries), BI can easily be chaotic, with random initialization avoiding stable orbits.

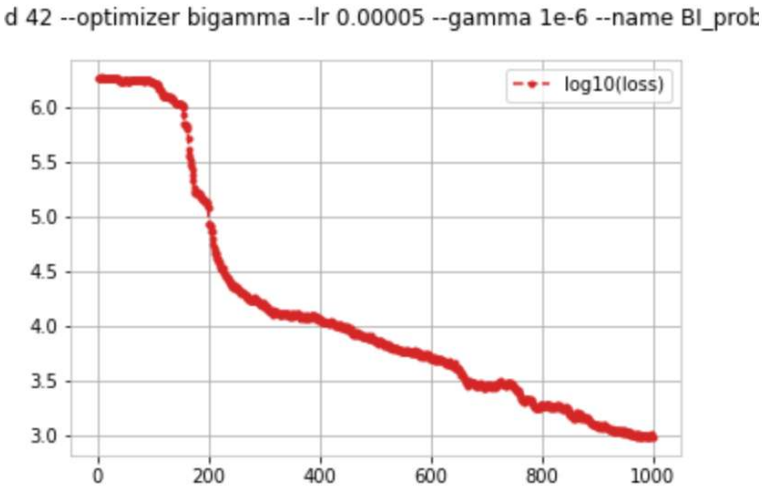
But if a particular problem (NN & Loss function) leads to long-lived orbits, we can add extra features to the algorithm (as in chaotic billiards problems) to stimulate faster mixing

Toy Example: $-\nabla^2 u + u^2 = f,$ $f = \frac{1}{8}\left(3 - 4(1 + 6400(x_1^2 + x_2^2))\cos(40(x_1^2 + x_2^2)) + \cos(80(x_1^2 + x_2^2)) - 640\sin(40(x_1^2 + x_2^2))\right)$

Original problem (stuck in orbit):

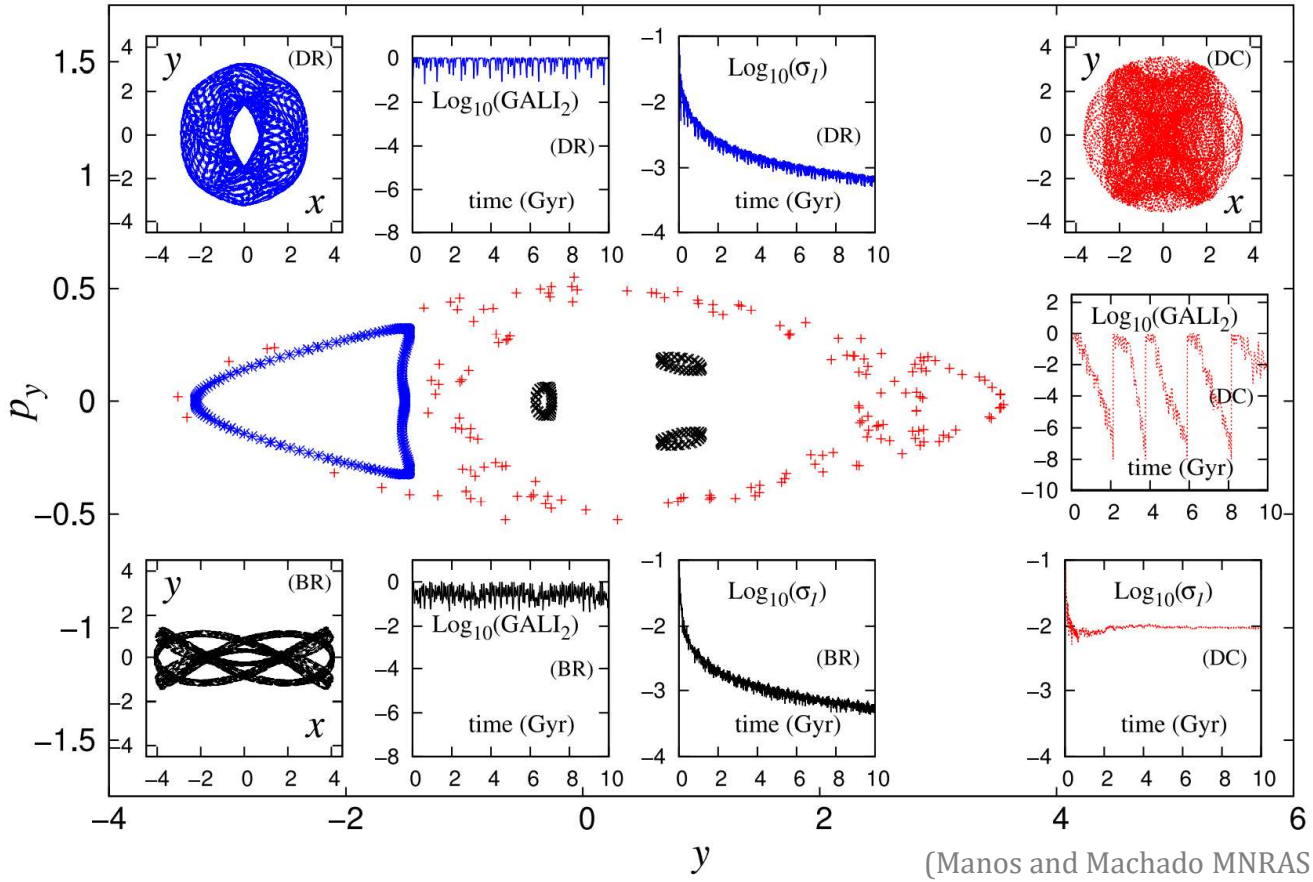


With added feature (unstuck):



Our redshifted BI dynamics is a bit like galactic dynamics, solar system, … where chaos (as well as long lived orbits) is familiar.

We add elements aimed at ensuring rapid mixing.



(Manos and Machado MNRAS '14).

Figure 5. The Poincaré Surface of Section defined by $x = 0$, $p_x \geq 0$ with $H = -0.19$, for three typical orbits (two regular and one chaotic) being integrated for 10 Gyr. The set of parameters for the bar, disc and halo components are chosen from the fits with the 3-d.o.f. TD Hamiltonian at $t = 7.0$ Gyr of the N-body simulation. In the insets, we depict their projection on the (x, y)-plane together with the GALI_2 and MLE σ_1 evolution in time (see Table 1 for the exact parameters and text for more details on these trajectories).

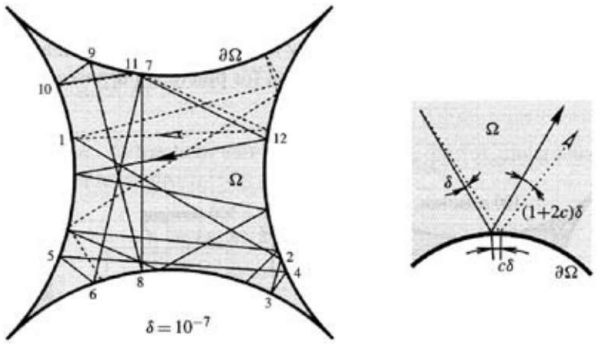
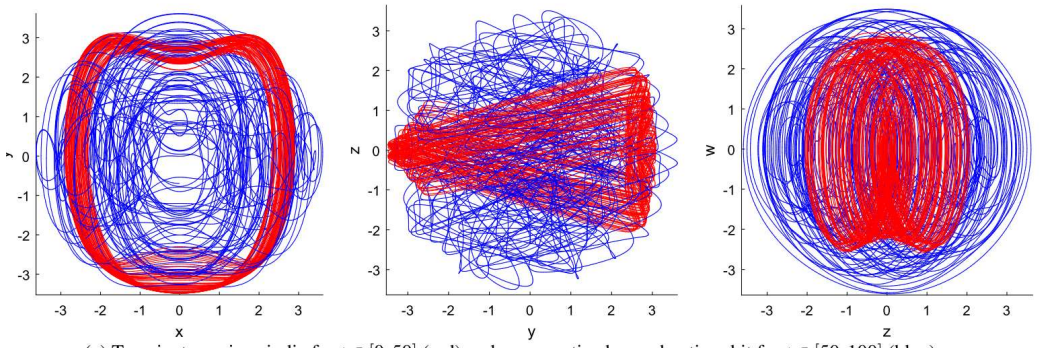


Figure 2. Illustration of the trajectory sensitivity to the initial conditions in a billiard model with convex borders.



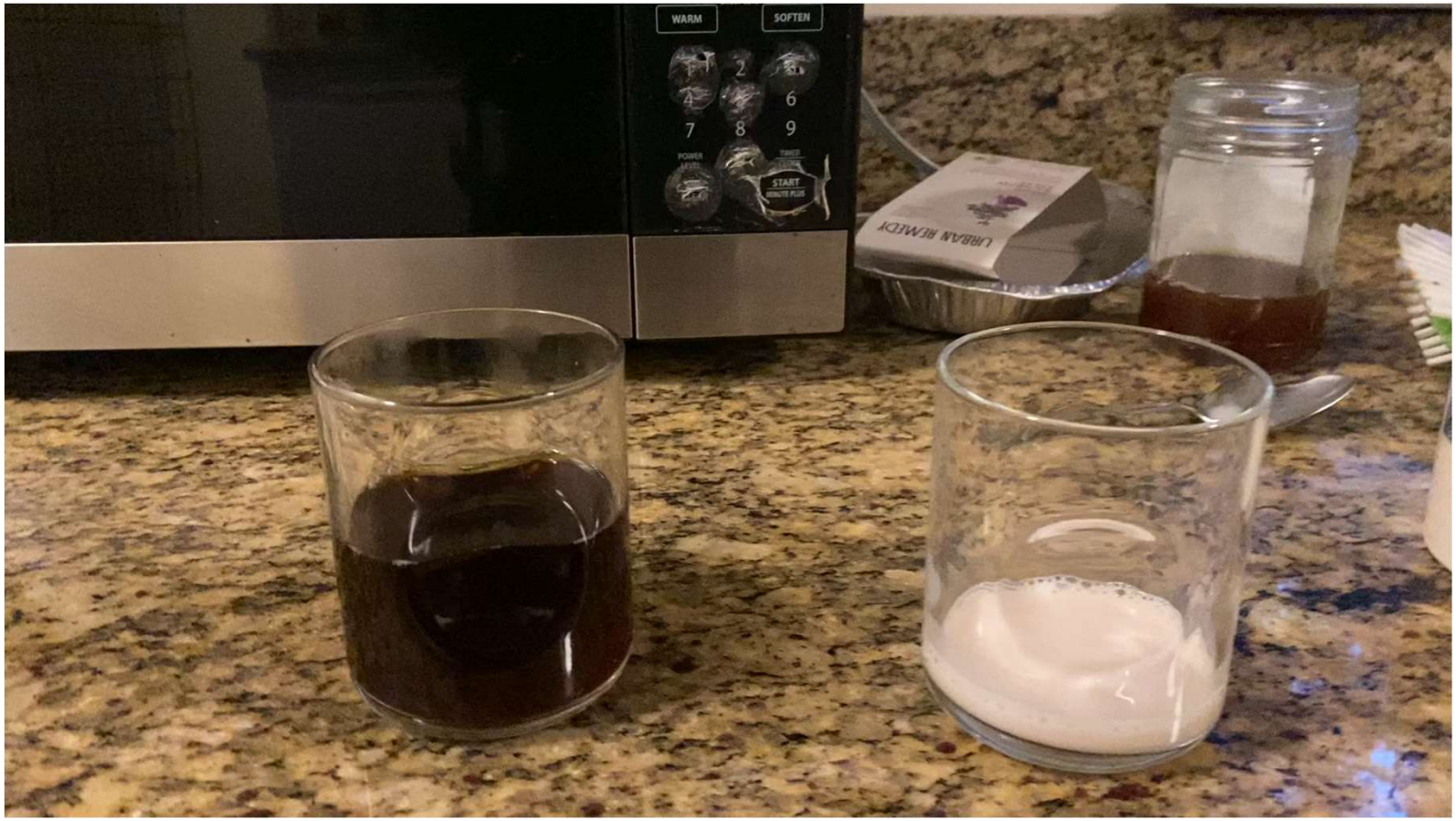
(a) Transient quasi-periodic for $t \in [0, 50]$ (red) and conservative hyperchaotic orbit for $t \in [50, 100]$ (blue);

Adding dispersing elements, (e.g. billiards or negative curvature) supports mixing (decay of correlations)

After some time, for a particle p in a droplet and phase space region R,

$$Prob(p \in R) \propto Vol(R)$$

(>ergodicity: $\langle f \rangle_t = \langle f \rangle_{phase\ space}$)


WARM
SOFTEN
START
URBAN REMEDY

BI algorithm:

Underlying discrete dynamics:

$$\theta_i(t+\Delta t) - \theta_i(t) = \Delta t\ \pi_i(t+\Delta t)\frac{V(\mathbf{\Theta}(t))}{E}$$

$$\pi_i(t+\Delta t) - \pi_i(t) = -\Delta t\frac{\partial_i V(\mathbf{\Theta}(t))}{2}(\frac{E}{V}+\frac{V}{E})$$

$$\sqrt{V(V+\vec{\pi}^2)} \equiv E$$

Plus:

- Initialization: option for E > V(t=0)
- E conservation enforced throughout (by rescaling of Π)
- Option: not enough progress down V => bounces:

$$\Pi \rightarrow (Random\ Rotation) * \Pi$$

- Option: user defined intervals => bounces regardless of progress (to help trajectories rapidly mix)

Measure in different regions gives predicted distribution over all solutions (given mixing):

V<< E:

$$Vol(\mathcal{M}_{\mathcal{I}}) = \frac{2\pi^{n/2}}{\Gamma(n/2)} E^{n-1} \int d^n(\theta - \theta_I) V^{-n/2}$$

Near minima:

$$V \simeq V_I + \frac{1}{2}\sum_{i=1}^{n} m_{Ii}^2(\theta_i - \theta_{Ii})^2 \qquad \eta_{iI} = m_{iI}(\theta_i - \theta_{Ii}) \Rightarrow V \simeq V_I + \frac{1}{2}\sum_{i=1}^{n} \eta_{Ii}^2$$

$$Vol(\mathcal{M}_{\mathcal{I}}) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \frac{E^{n-1}}{\prod_i m_{Ii}} \int d^n\eta V^{-n/2} = \left(\frac{2\pi^{n/2}}{\Gamma(n/2)}\right)^2 \frac{E^{n-1}}{\prod_i m_{Ii}} \int d\eta \frac{\eta^{n-1}}{(V_I + \frac{1}{2}\eta^2)^{n/2}}$$

$$Vol(\mathcal{M}_{\mathcal{I}}) \to b_n \left(\frac{2\pi^{n/2}}{\Gamma(n/2)}\right)^2 \frac{E^{n-1}}{\prod_i m_{Ii}} \log(V_I) \qquad V_I \to 0, \quad \text{fixed n}$$

We can check this distribution using our discretized algorithm:

$$V = -\exp\left(-0.4|x-x_1|^2\right) - (1-\epsilon)\exp\left(-0.8|x-x_2|^2\right) + 10^{-3}|x-x_1|^2|x-x_2|^2 + 1 \tag{2}$$

Theory:

$$\frac{\mathrm{Vol}(\mathcal{M}_1)}{\mathrm{Vol}(\mathcal{M}_2)} = \frac{e_2}{e_1} \sim 1.93$$

Experiment:

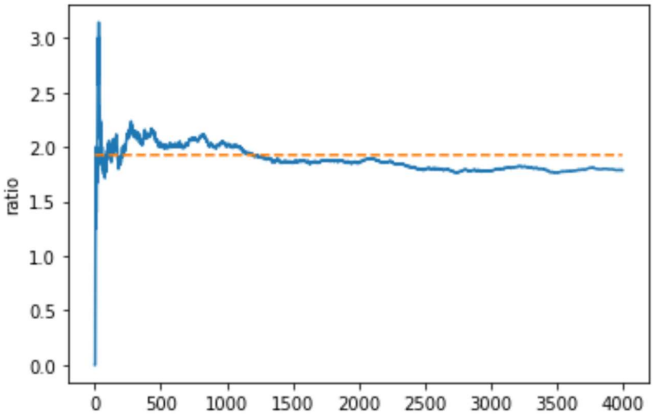
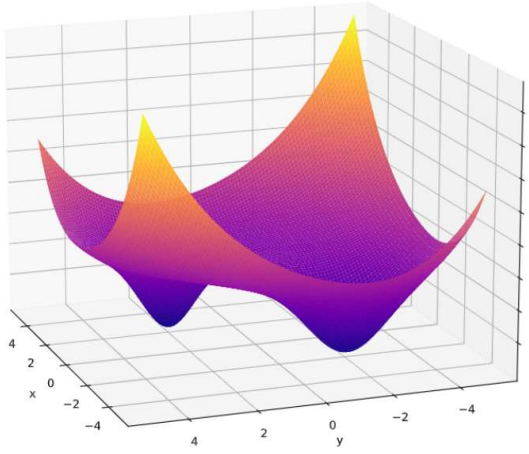
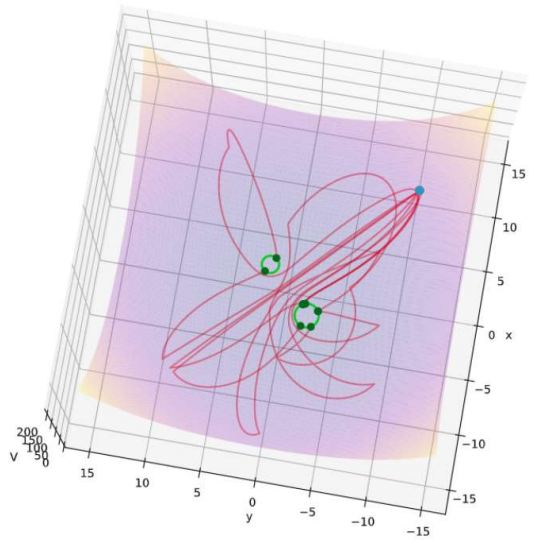


Figure 4: Partial ratios.



(a) The potential with two basins of Eq. (2)

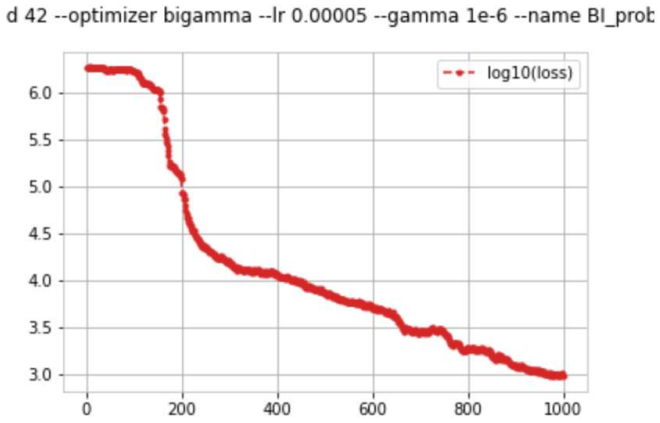


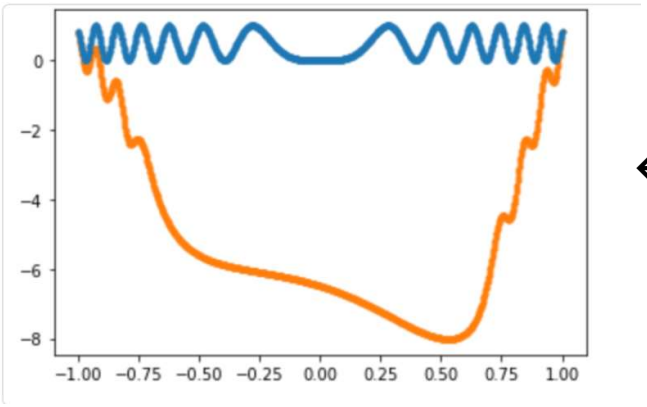
(b) A small sample of trajectories

Agreement within 10%.

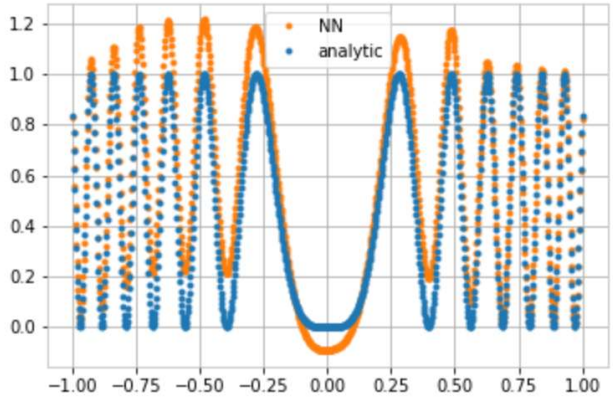
On our PDE, the BI optimizer solves the PDE (finding multiple solutions)

BI





←1d slice of domain→

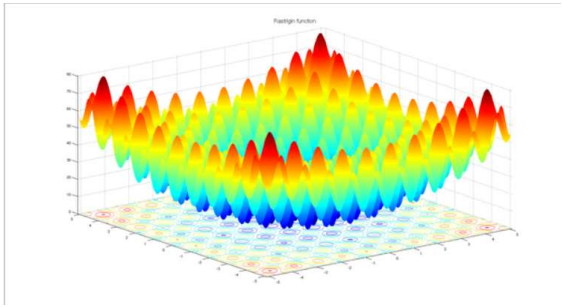


BI (and other frictionless, E conserving dynamics)	SGD-momentum (and anything friction-based)
Conservative Hamiltonian Dynamics	Friction => contraction of phase space
Cannot get stuck in local minimum	Can contract to local minimum
Cannot overshoot V=0	Can overshoot V=0
Evolution on shallow region: $\|\dot{\theta}\| \sim \sqrt{\|\{(\theta - \theta_0) \cdot \partial V\}\|}$	Evolution on shallow region slower: $\|\dot{\theta}\| \sim \|\frac{\partial V}{friction}\|$
Analytic prediction for distribution among multiple solutions (given mixing), with same initialization	Intuition that can cover multiple solutions via distribution of initializations + stochasticity
Can choose V_0 to stop at any value of loss	`early stopping' sometimes used

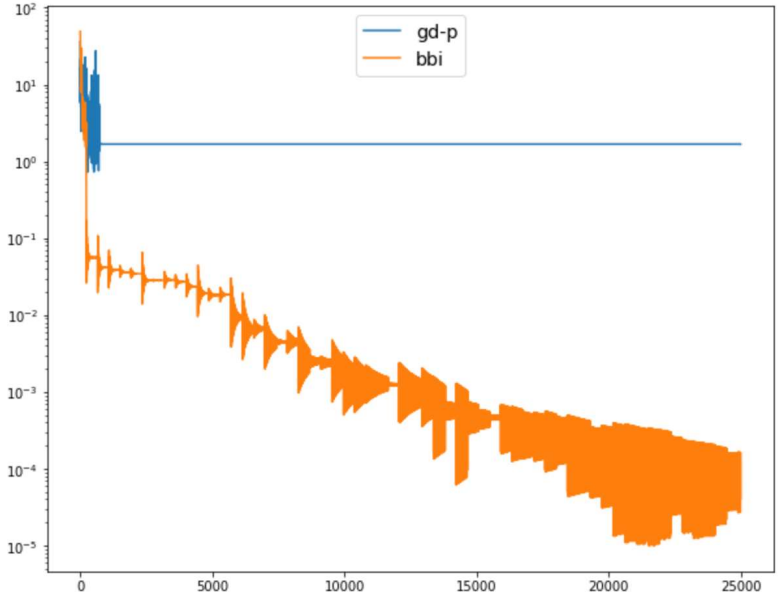
These statements persist with noise (mini-batches) in our prescription, more below…

Rastrigin function
(non-convex test function for optimization)

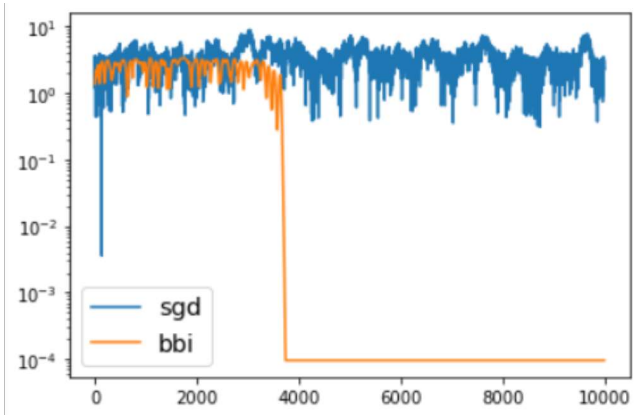
$$f(\mathbf{x}) = An + \sum_{i=1}^{n} \left[x_i^2 - A\cos(2\pi x_i) \right]$$



n=5, following hyper-parameter optimization

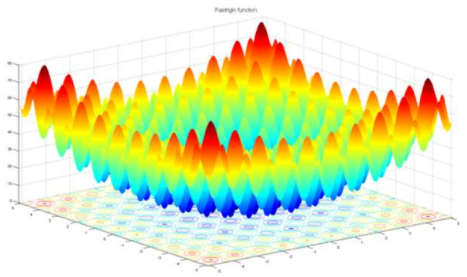


GD may be helped by `catapult' mechanism Lewkowycz et al '20, . But it appears less predictable (bounces out of the basin of the minimum:

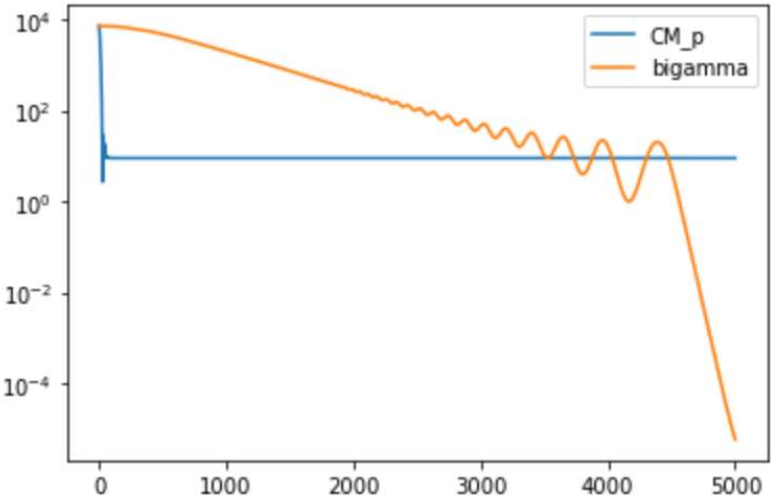


400-dimensional Rastrigin function + ϵx^8 (non-convex test function for optimization)

$$f(\mathbf{x}) = An + \sum_{i=1}^{n} \left[x_i^2 - A\cos(2\pi x_i) \right]$$

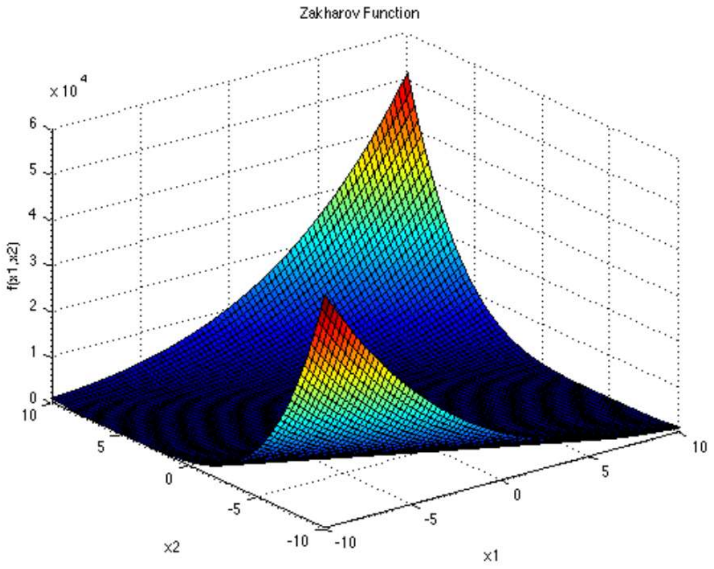


```
run experiment-Rastrigin-BI-CMP.py

100%|██████████| 1000/1000 [00:12<00:00, 77.85trial/s, best loss: 6904.38975291786]
100%|██████████| 1000/1000 [00:09<00:00, 106.06trial/s, best loss: 2.3404389537518e-08]

Best parameters
CM_p: {'gamma': 1.8946762796055006, 'stepsize': 0.07000604163714612}
bigamma: {'gamma': 2.5488927707048213e-06, 'stepsize': 0.0009999976086160324}
```

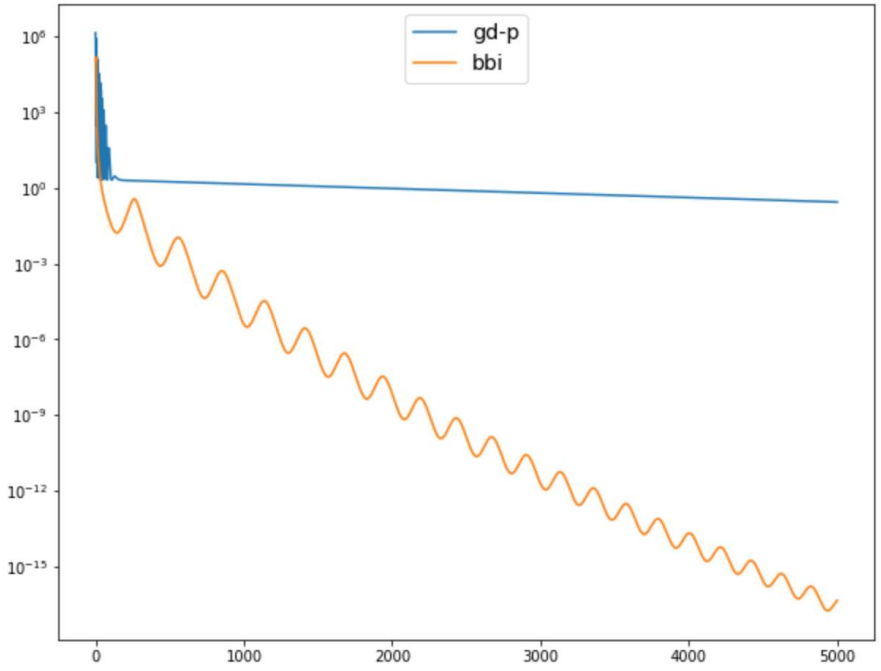


Zakharov function (benchmark): shallow valley

$$f(\mathbf{x}) = \sum_{i=1}^{d} x_i^2 + \left(\sum_{i=1}^{d} 0.5 i x_i\right)^2 + \left(\sum_{i=1}^{d} 0.5 i x_i\right)^4$$

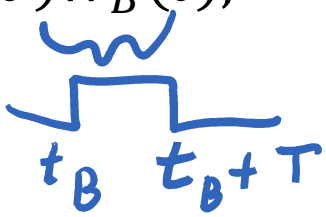


d=10 Zakharov



Noisy case (mini-batches):

$$V(\theta(t),t) = \sum_B V^B(\{x\}_B,\theta) W_B(t), \qquad V_{full} = \sum_{\{x\}_B} V^B \quad e.g.\ V^B > 0\ \forall B$$



Time dependent potential (nonetheless we renormalize to the original E).
One can think of a given batch trajectory as deterministic.
Retains the main features:

- Cannot stop at local minimum (V>0)
- Will stop at global minimum due to speed limit

Also interesting to study ensemble averages, generalized Brownian motion:

Discrete Fluctuation-Dissipation relations generalizing Yaida '18 (SGD+momentum)

$$\langle [|\frac{\partial_i V^B}{V^B}\theta_j|] \rangle = \langle [|\frac{\partial_j V^B}{V^B}\theta_i|] \rangle$$

$$\langle\ V(\theta_i \Pi_j + \theta_j \Pi_i)\ \rangle = \Delta t E\ \langle\ \frac{1}{4}(\partial_i V \theta_j + \partial_j V \theta_i) - [|\frac{(V^B)^2}{2E^2}\Pi_i \Pi_j|]\ \rangle$$

Careful continuum limit with noise:

rescaling for fixed E

$$\ddot{\theta}_i = -\frac{1}{2}\partial_i V + \dot{\theta}_i(\frac{\dot{\mathbf{\Theta}} \cdot \nabla V}{V}) + \dot{\theta}_i \sum_B \delta(t - t_B)(1 - \lambda(\Delta V^B) + \frac{\Delta V^B}{V^B})$$

$$+\text{bounces} + \mathcal{O}(\Delta t)$$

$1 - \lambda(\Delta V_B)$ is of order $\Delta V_B / V_B$ when this ratio is small.

No friction term

Contrast to SGD+momentum:

e.g. Kunin Sagastuy-Brena, Gillespie, Tanaka, Ganguli, Yamins '21

$$\frac{\Delta t}{2}(1+\beta)\ddot{\theta} + (1-\beta)\dot{\theta} = -\partial V^B$$

$$v_{k+1} - \beta v_k = -\partial V_k, \qquad \theta_{k+1} - \theta_k = \Delta t\ v_{k+1}$$

Late-time Brownian motion (preliminary comparison):

Normally (≈somewhat like in SGD-momentum): $\frac{\mathrm{d}\langle\theta^2\rangle}{dt} \propto \langle\dot{\theta}^2\rangle$

BI: ...+ $\mathrm{d}\frac{\langle\theta^2\rangle}{dt} \sim \langle\dot{\theta}^2\rangle < V$

BI explores the landscape in a very different way, with or without noise.

Distinctive behavior vis a vis local and global minima persists with noise.

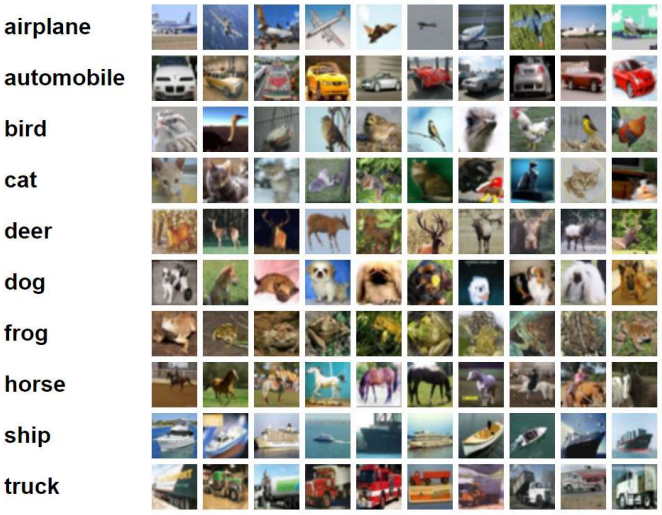
Cifar image data set (all optimizers work)

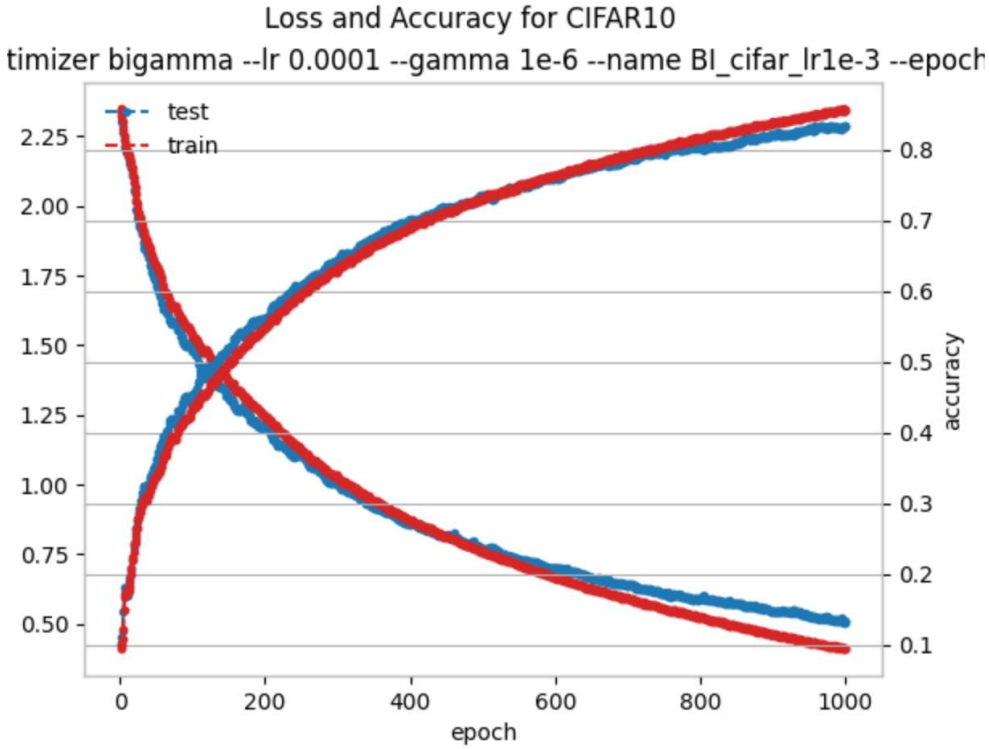
The CIFAR-10 dataset

The CIFAR-10 dataset consists of 60000 32x32 colour images in 10 classes, w

The dataset is divided into five training batches and one test batch, each with 1
training batches contain the remaining images in random order, but some traini
contain exactly 5000 images from each class.

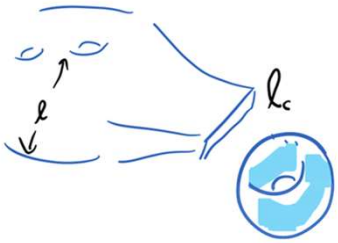
Here are the classes in the dataset, as well as 10 random images from each:



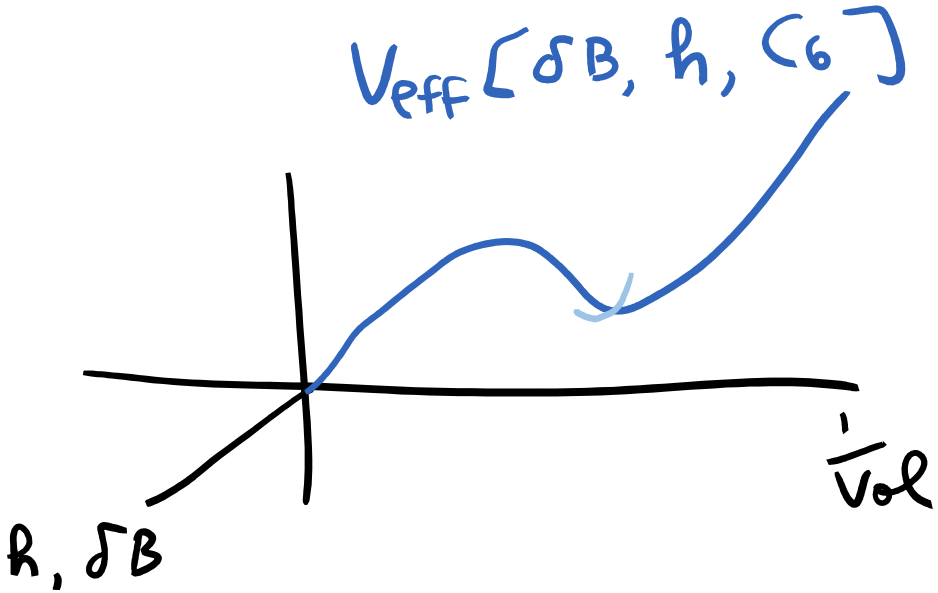


Application to PDEs in new mechanism for Λ from string theory

(w/G.B. De Luca, G. Torroba '21):



M theory (EFT: 11d SUGRA) on finite-volume hyperbolic space with small systole, automatically-generated Casimir energy, 7-form flux yields immediate volume stabilization.



Strong positive Hessian contributions from **hyperbolic rigidity** and from **warping** (redshifting) effects on conformal factor and on Casimir energy.

4d effective potential

Douglas '09

net curvature term

$\ell_{11}^9 \rho_c(R_c) \sim -\frac{\ell_{11}^9}{R_c^{11}}$

$$V_{eff}[g^{(7)}, C_6] = \frac{\ell_{11}^9}{2G_N^2} \frac{\int d^7y \sqrt{g^{(7)}} u^2|_c \left([-R^{(7)} - 3\left(\frac{\nabla u}{u}\right)^2 \Big|_c] - \frac{1}{4}\ell_{11}^9 T^{(Cas)\mu}_{\mu} + \frac{1}{2}|F_7|^2 \right)}{(\int d^7y \sqrt{g^{(7)}} u|_c)^2}$$

$$ds^2 = e^{2A(y)} ds^2_{dS_4} + e^{2B(y)} (g_{\mathbb{H}ij} + h_{ij}) dy^i dy^j \qquad u(y) = e^{2A(y)}$$

u(y) satisfies GR constraint (its equation of motion):

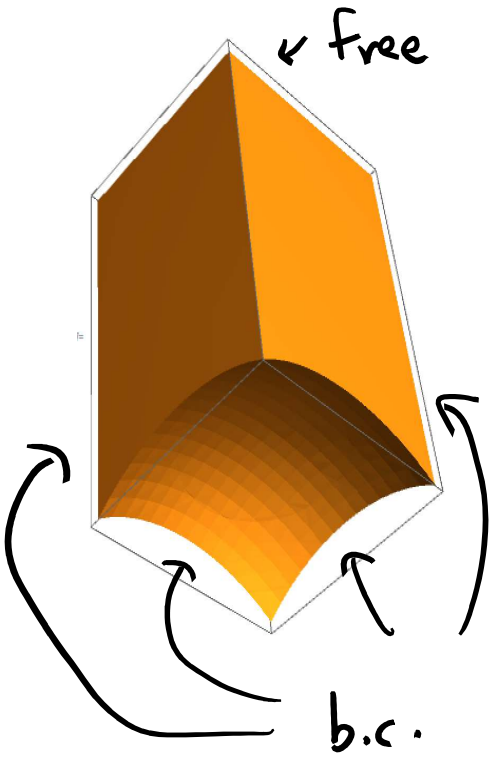
$$\left(-\nabla^2 - \frac{1}{3}\left(-R^{(7)} - \frac{1}{4}\ell_{11}^9 T^{(\mathrm{Cas})\mu}_{\mu} + \frac{1}{2}|F_7|^2 \right) \right) u = -\frac{C}{6}$$

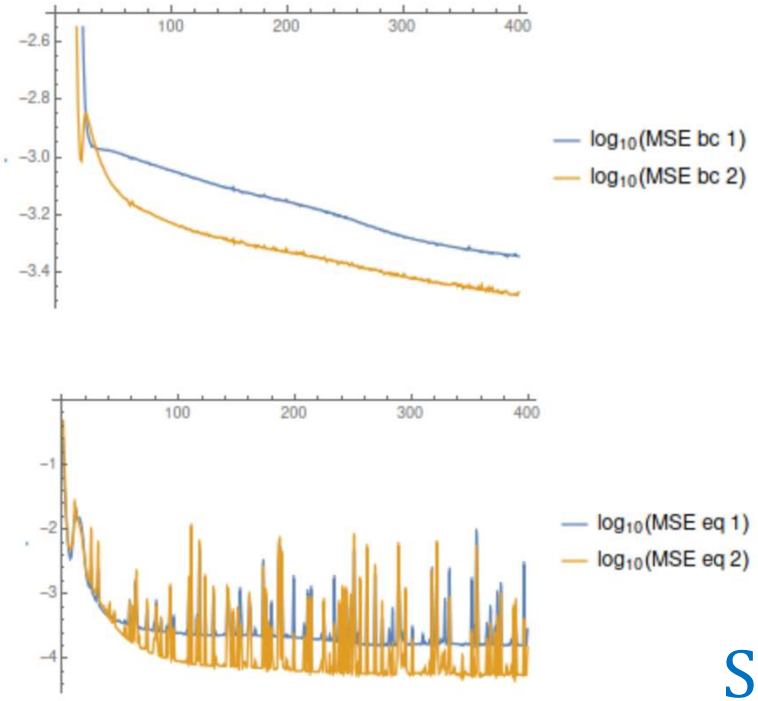
Like a Schrodinger problem for

$$C\ell^2 \sim H^2\ell^2 \ll 1$$

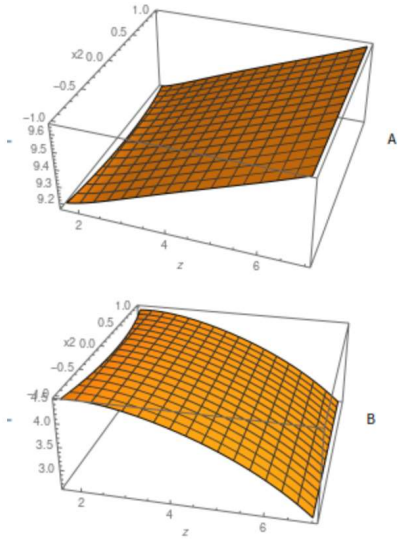
$$\Rightarrow \quad V_{eff} = \frac{C}{4G_N} = \frac{R^{(4)}_{\mathrm{symm}}}{4G_N}\,.$$

H_3 warmup example:





Loss



Slice of approximate solution for warp and conformal factors;

Numerical study of this class of compactifications is fully specified and well-posed, including the stress-energy sources relevant for dS:

- H_7/Γ explicit projection of H_7, can also be constructed as gluing of explicit set of polygons.
- $\Gamma \Rightarrow$Casimir energy
- F_7 solution explicit in terms of metric
- Parametric limit(s) involving covers and filled cusps to compare to.

For ML, can consider PDE's, V_{eff}, or slow roll functionals ϵ_V, η_V as natural loss functions to explore.

Summary:

BI for AI
(et al)

$$S = -\int Loss(\vec{\theta})\sqrt{1 - \frac{\dot{\vec{\theta}}^2}{Loss(\vec{\theta})}}$$

$\vec{\theta} = \{W, b\}$

- Energy-conserving dynamics (no friction), yet slows at Loss → 0 , cannot overshoot V=0, cannot get stuck in local minimum

- If mixing (as well as ergodic), spends large fraction of time near $\dot{\theta}^2 \simeq Loss \simeq 0$ in phase space and captures multiple solutions

So far, spent few resources (in defining and testing the algorithm and its agreement with theory).
Future plans: apply to scientific ML (e.g. large-scale structure,…,protein folding), higher-dimensional PDEs (including landscape, Veff etc), other data sets.