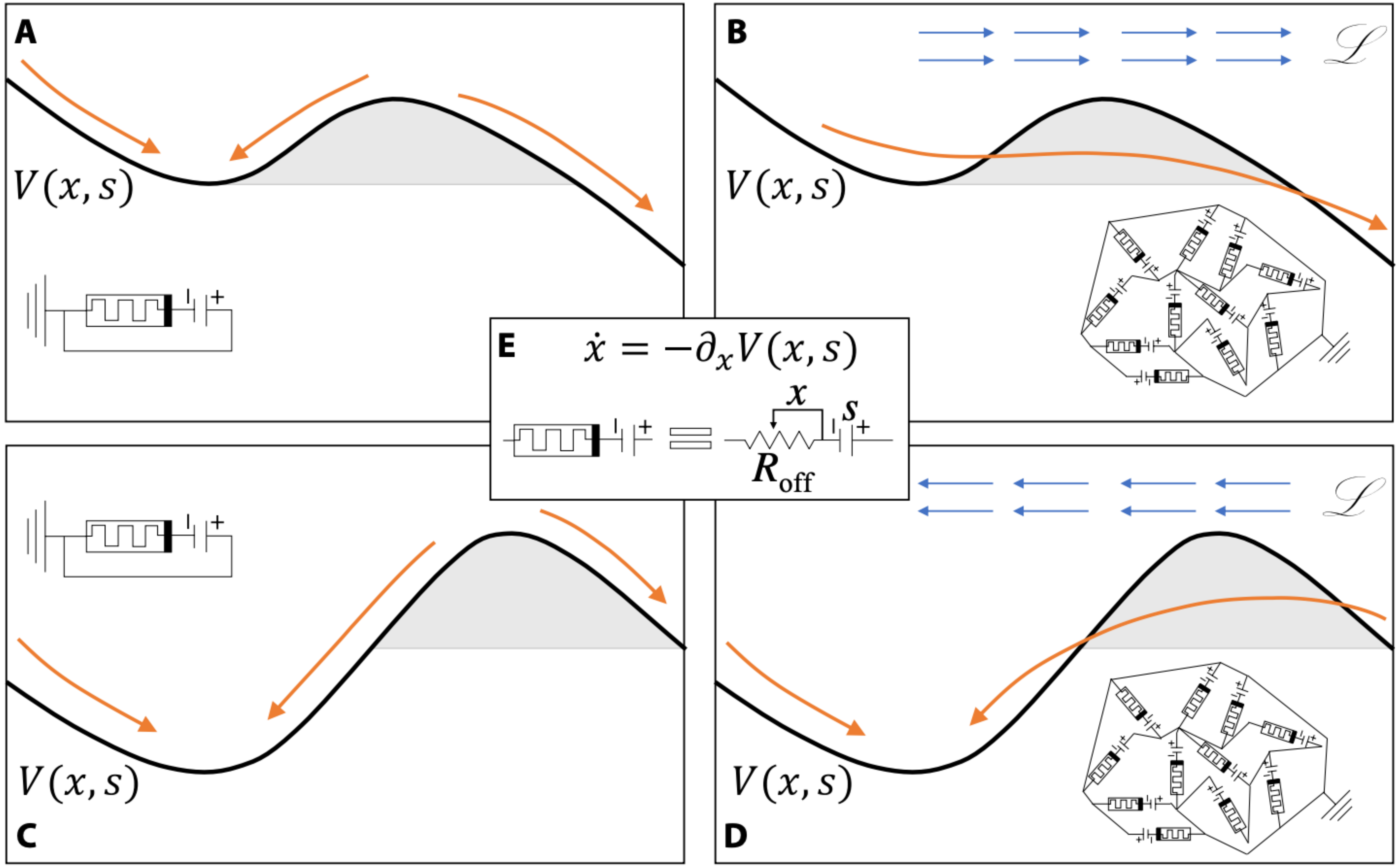
On the global minimum convergence of non-convex deterministic functions via Stochastic Approximation

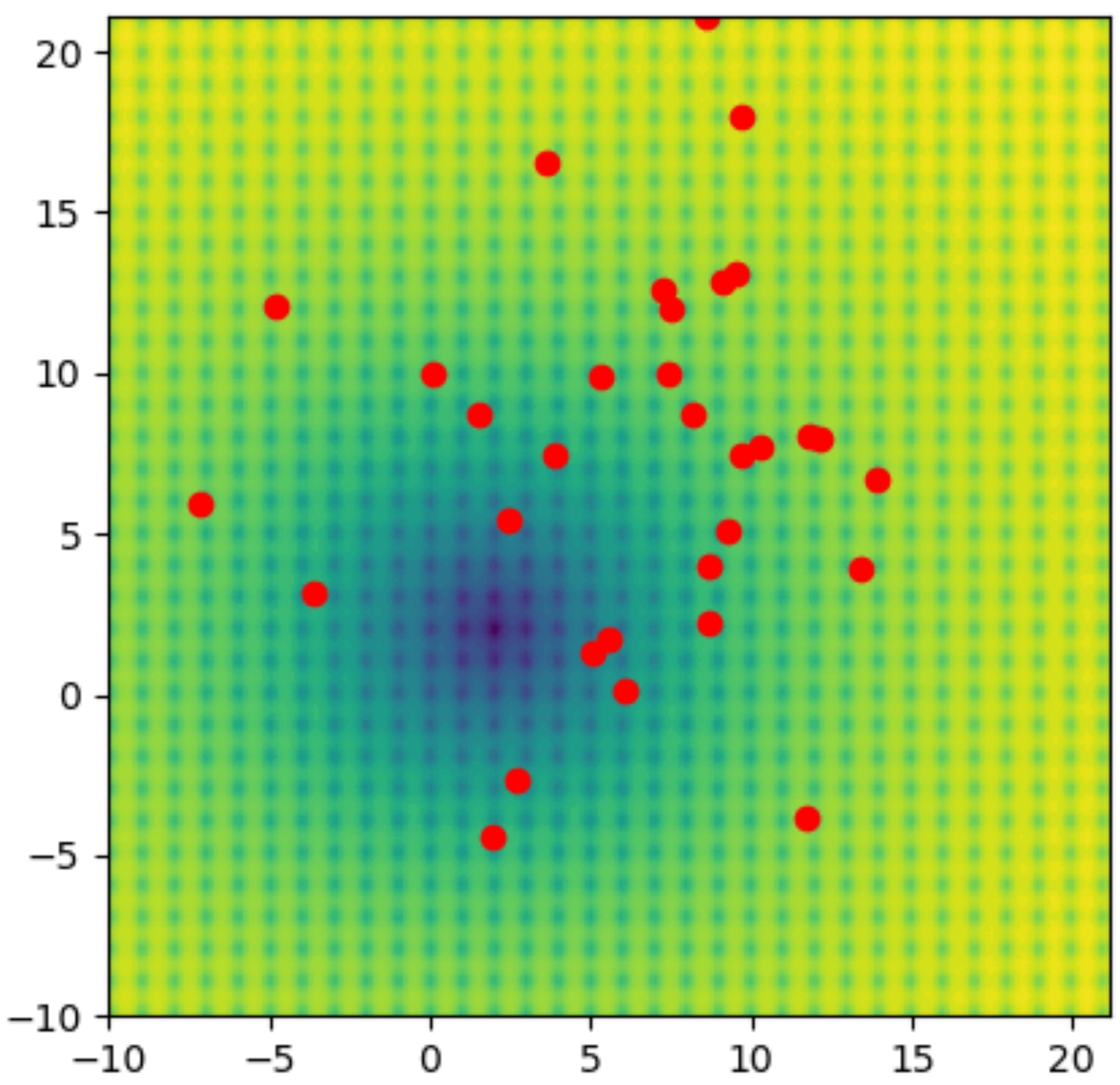
Charlie Chen (mentors: Prof. Stefano Martiniani and Dr. Guanming Zhang)
July 27



(Caravelli et al. 2021)

Quick demonstration of restart strategy

On Ackley function



Content

- Projective Embeddings of Dynamical Systems (PEDS)
- PEDS as particle interactions
- Inspiration for SA-PEDS
- Algorithm: SA-PEDS
- Intuitions for SA-PEDS
- Experiments
- Discussions

Projective Embeddings of Dynamical System (PEDS)

(Caravelli et al. 2023)

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \to \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & \cdots & & x_{N,m} \end{pmatrix} \quad Y_m$$

- The optimization problem is: $\min_X F(X)$, where $X \in \mathbb{R}^m$.
- Extend the variable to $M \in \mathbb{R}^{N\times m}$. Denote the column vector by $Y_j = M[:,j]$.
- The update for Y_j^t is then
  - $Y_j^{t+1} - Y_j^t = -\gamma(\Omega\ \Phi(\nabla F; Y_1^t, Y_2^t, \ldots, Y_m^t) + \alpha(I - \Omega)Y_j^t)$,
- where Ω is a projection matrix, i.e. $\Omega^2 = \Omega$, Φ is called **matrix map**, γ is the learning rate, and α is some hyper parameter.

PEDS as particle interactions

For a particular case in PEDS

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \to \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & \cdots & & x_{N,m} \end{pmatrix}$$

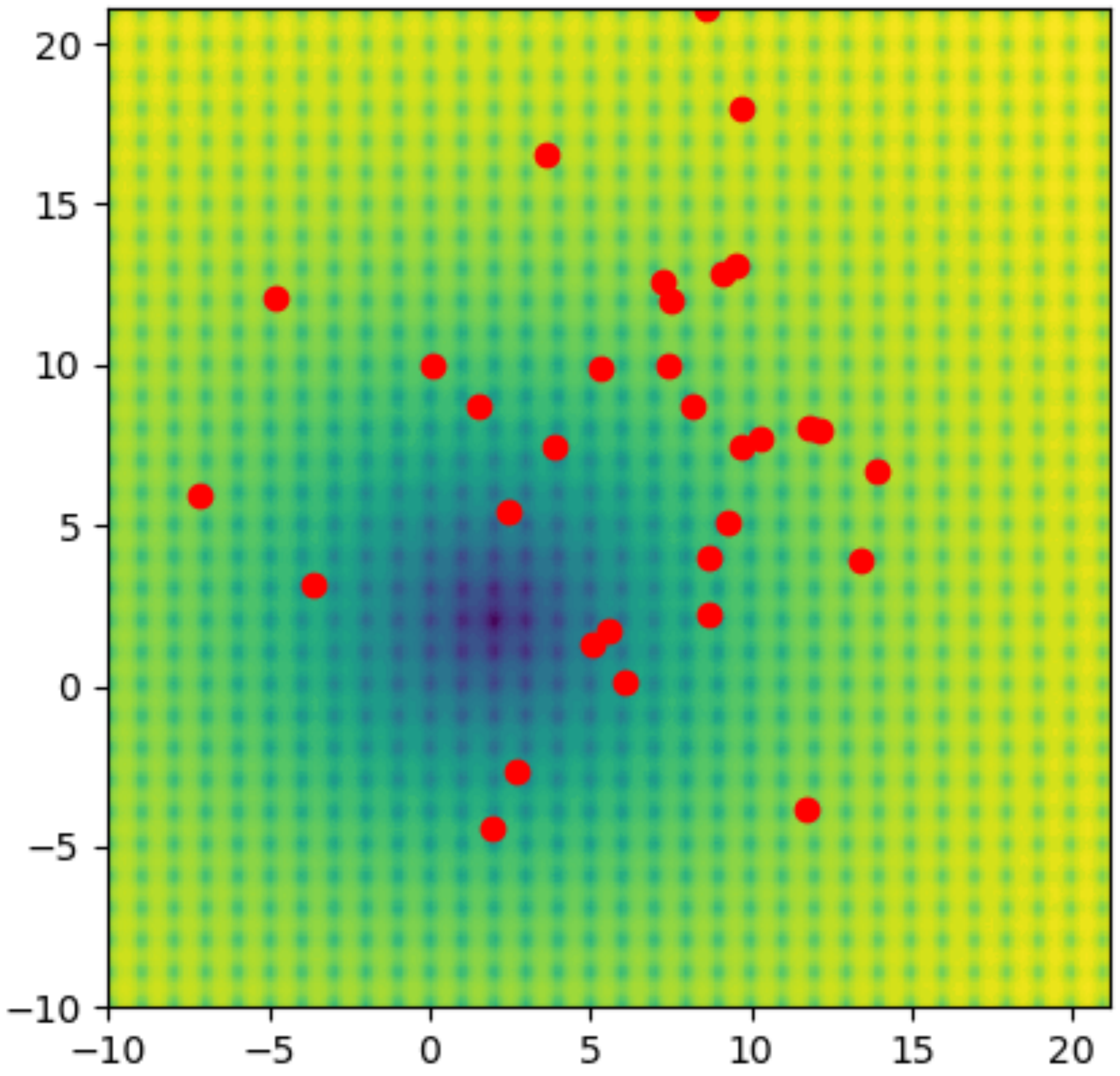
R_N

X_m

- $Y_j^{t+1} - Y_j^t = -\gamma(\Omega\, \Phi(\nabla F; Y_1^t, Y_2^t, \ldots, Y_m^t) + \alpha(I - \Omega)Y_j^t)$, for $j = 1,\ldots,m$
- For a particular choice of Ω and Φ, it can be shown that the update is equivalent to (see write-up for details)
- $$R_i^{t+1} - R_i^t = -\gamma\left(\frac{1}{N}\sum_{i=1}^{N} \nabla F(R_i^t) + \alpha(R_i^t - \overline{R}^t)\right)$$, for $i = 1,\ldots,N$,
- where R_i is the row vector of M and $\overline{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$, namely the center of mass.

Quick demonstration of PEDS

On Ackley function



Inspiration for SA-PEDS

How PEDS can be seen as a Stochastic Approximation algorithm

- $R_i^{t+1} - R_i^t = -\gamma \left(\frac{1}{N} \sum_{i=1}^{N} \nabla F(R_i^t) + \alpha(R_i^t - \overline{R}^t) \right)$
- Instead of treating R_i as deterministic, we treat it as samples from a distribution.
- For R_i be drawn from $\mathcal{N}(\theta, \sigma^2)$, the first term is the empirical approximation of $\mathbb{E}_{R\sim\mathcal{N}(\theta,\sigma)} \nabla F(R)$. Here, θ is the center of mass, similar to $\overline{R}$.
- The second term pulls all the particles to their center of mass, which is equivalent to decrease the variance of next samples, i.e. decrease σ.
- Stochastic Approximation Algorithm deals with $f(\theta) = \mathbb{E}_\xi F(\theta, \xi)$.

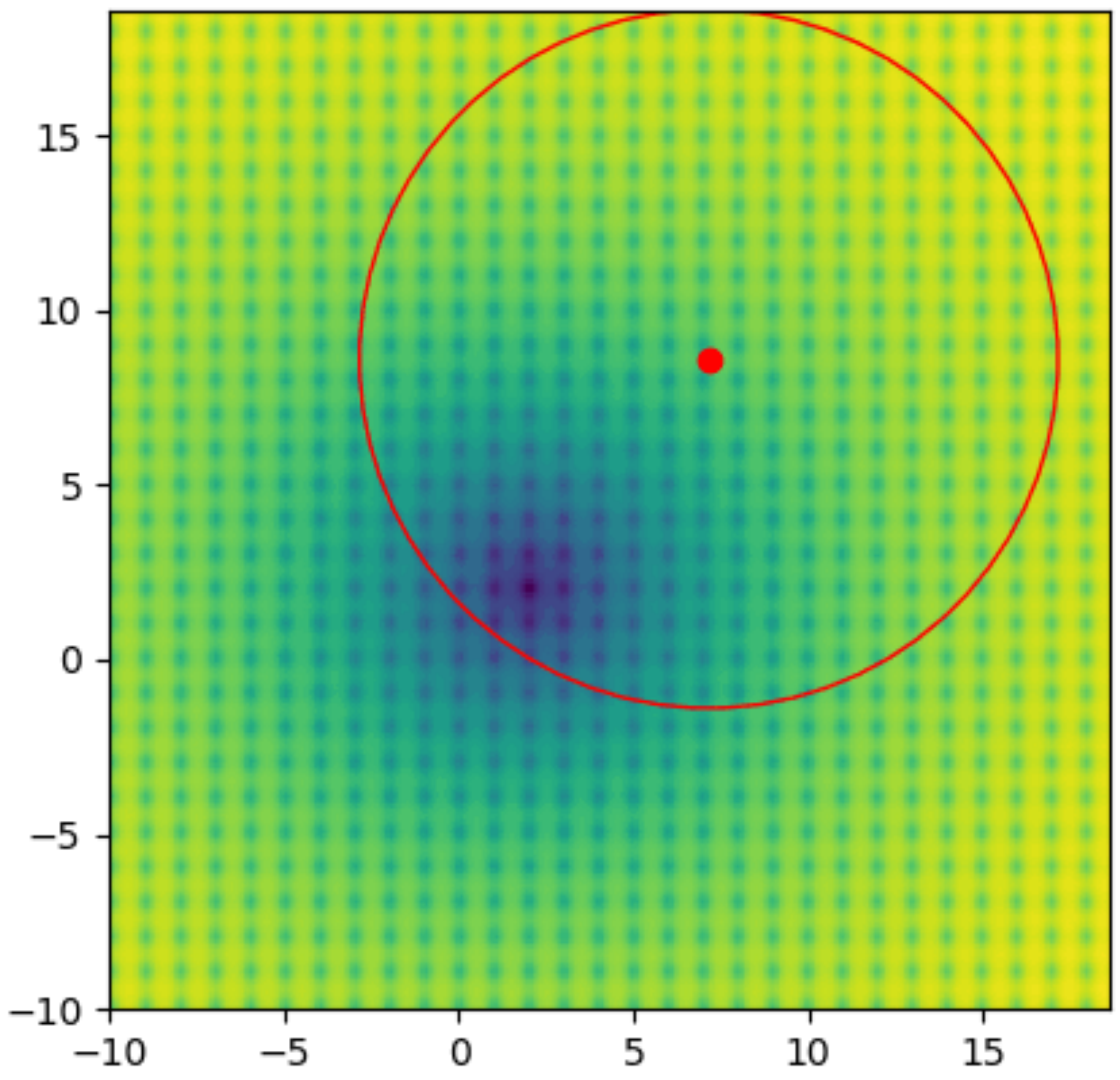
SA-PEDS

Stochastic Approximation Projective Embedding of Dynamical Systems

- Target: $\min_{\theta,\sigma} \mathbb{E}F(R)$, subject to $R \sim \mathcal{N}(\theta, \sigma)$.
- Given $\theta_0, \sigma_0, \gamma, \eta$
- For $t = 0,1,2,\ldots, T_{max}$ or stopping condition is met
    - Draw N samples $R_1^t, \ldots, R_N^t$ from $\mathcal{N}(\theta, \sigma^2)$.
    - Compute the gradient $g_t = \dfrac{1}{N}\sum_{i=1}^{N} \nabla F(R_j^t)$ and update $\theta_{t+1} = \mathrm{optim}(\theta_t, g_t, \gamma)$.
    - Shrink $\sigma_{t+1} = \max(\sigma_t - \alpha, 0)$, where α is some fixed parameter
- The last θ is our minimizer.

Quick demonstration of SA-PEDS

On Ackley function

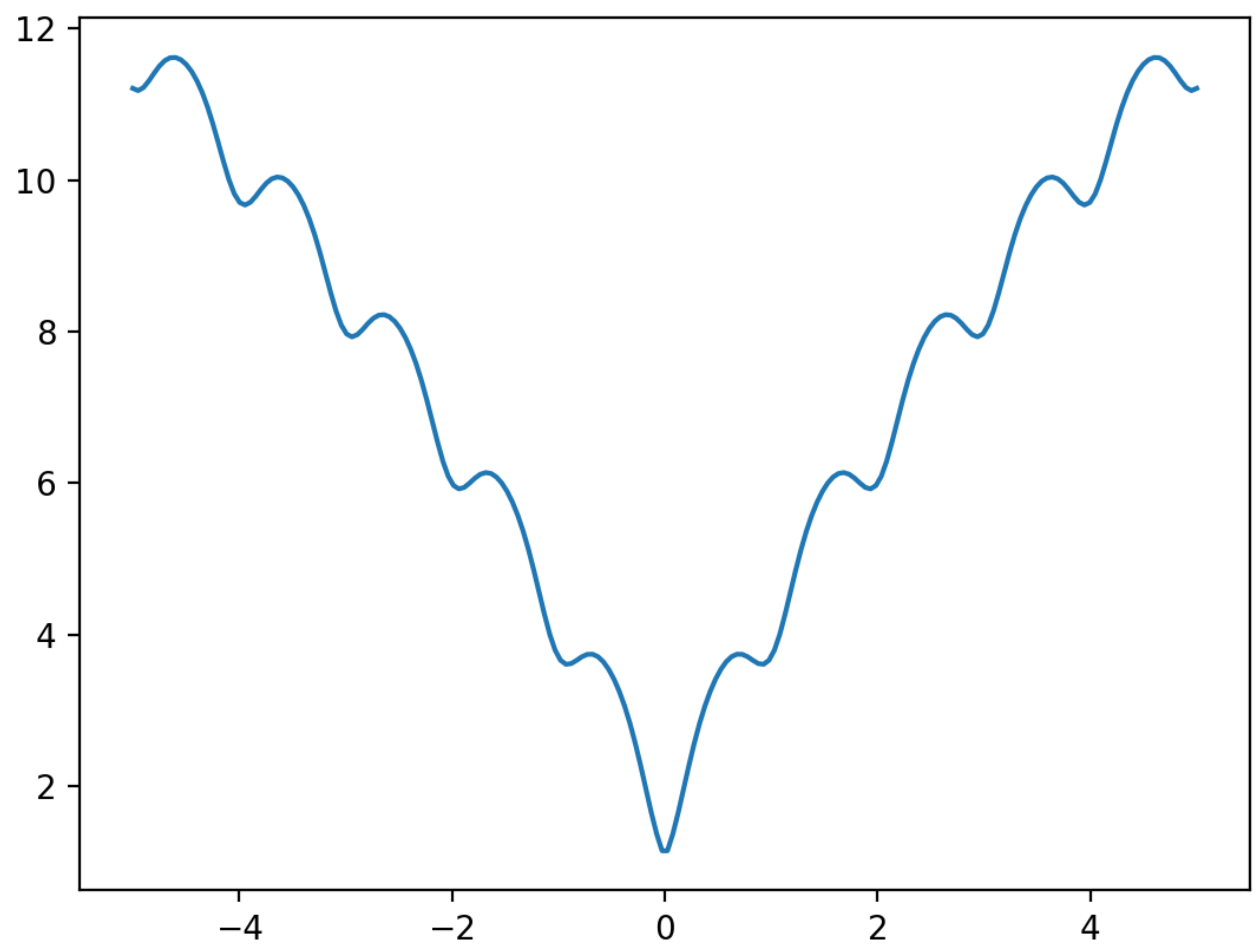


Intuitions for SA-PEDS

Why this methods can work?

- For $R \sim \mathcal{N}(\theta, \sigma)$, we have
- $$\mathbb{E}\nabla F(R) = \int \nabla F(R)\mathcal{N}(R; \theta, \sigma^2 I)dR = \int \nabla F(R)\rho(\theta - R)dX = \nabla F * \rho(\theta),$$
  - where $\rho(X) \approx e^{-\|X\|^2}$ (up to constants)
- This is as smooth as the Gaussian density function
- This is also called Randomized Smoothing, in the context of non-smooth Stochastic Gradient Descent (Duchi et al. 2012).

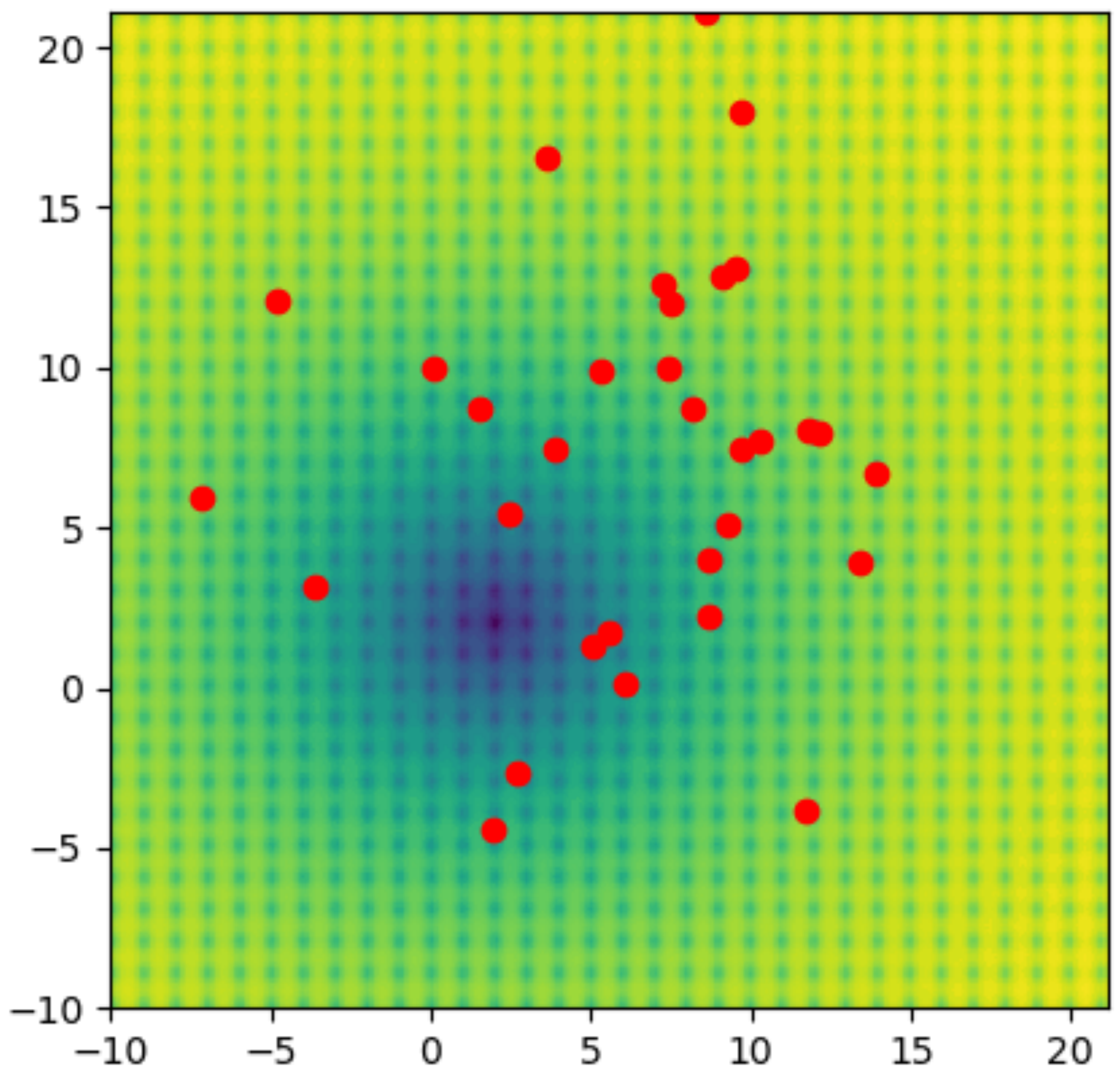
Experiments



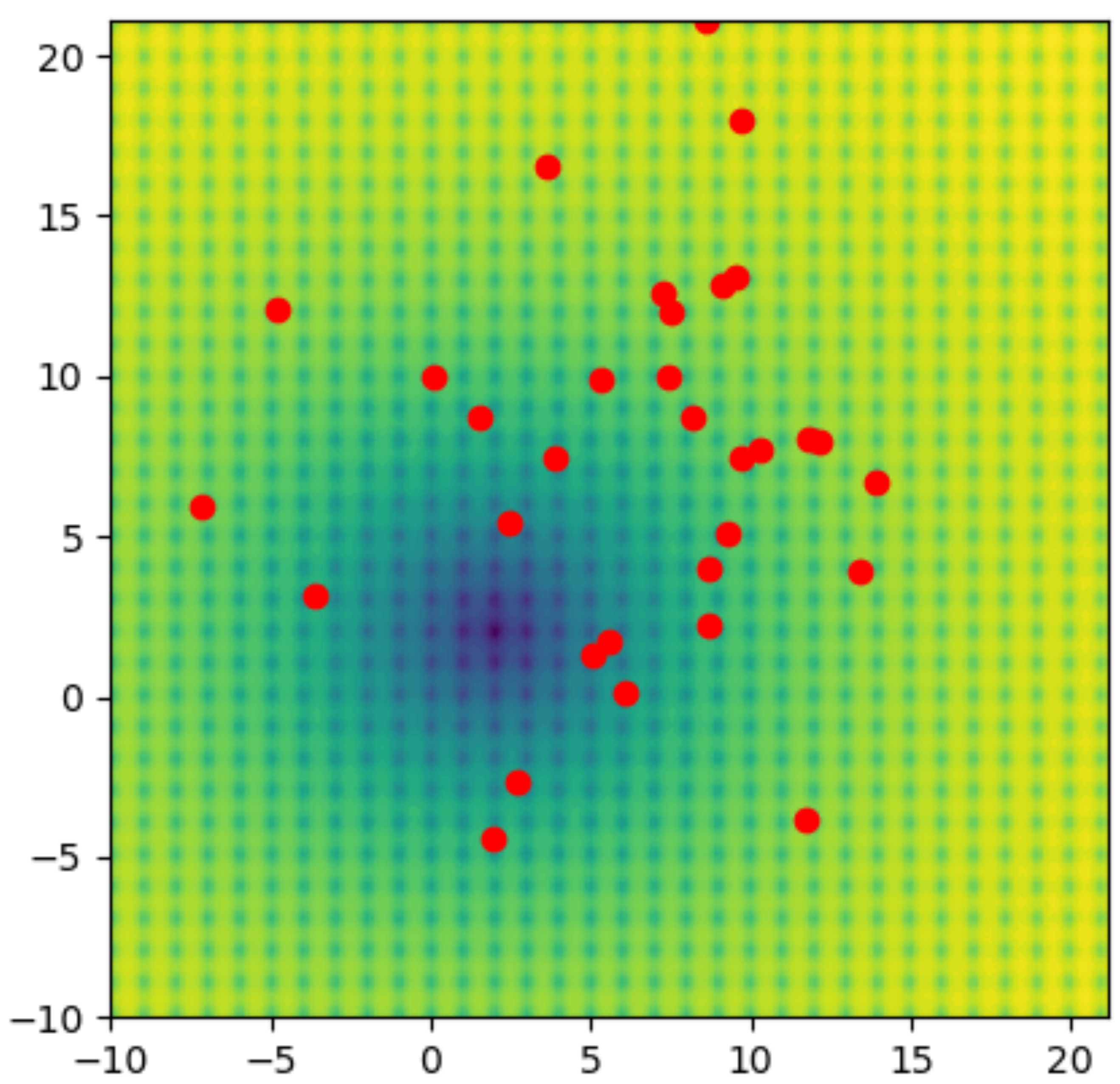
- Test function: Ackley function
- Approaches:
    - Restart: take different initial values and optimize.
    - PEDS: the original PEDS algorithm
    - SA-PEDS: the algorithm we proposed
- Interesting variables:
    - Success rate: if any particle finds the global min
    - Convergence time: how long does the convergence take

Code: https://github.com/charliezchen/SA-PEDS

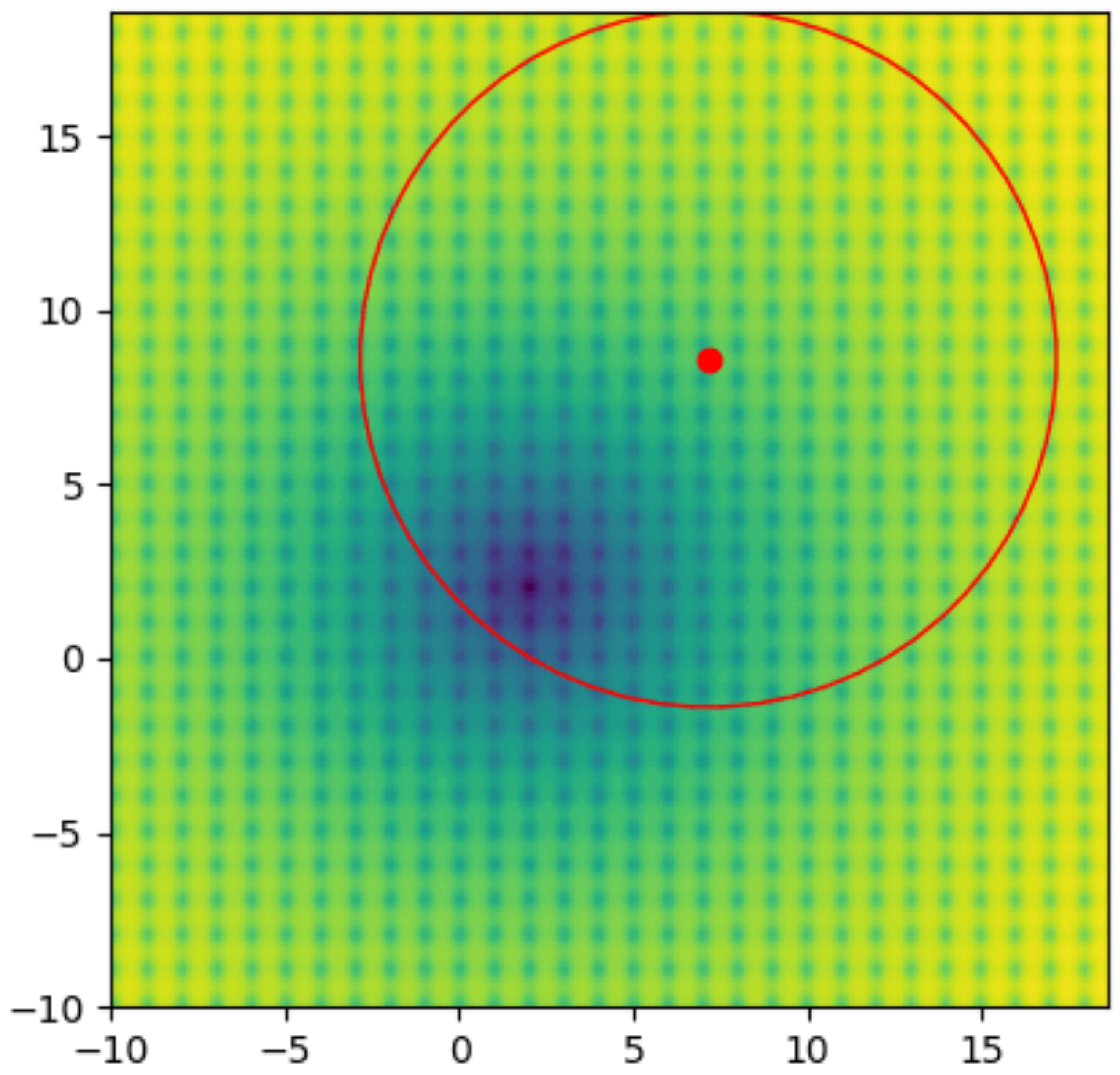
Experiments (m=2, N=20)



Restart



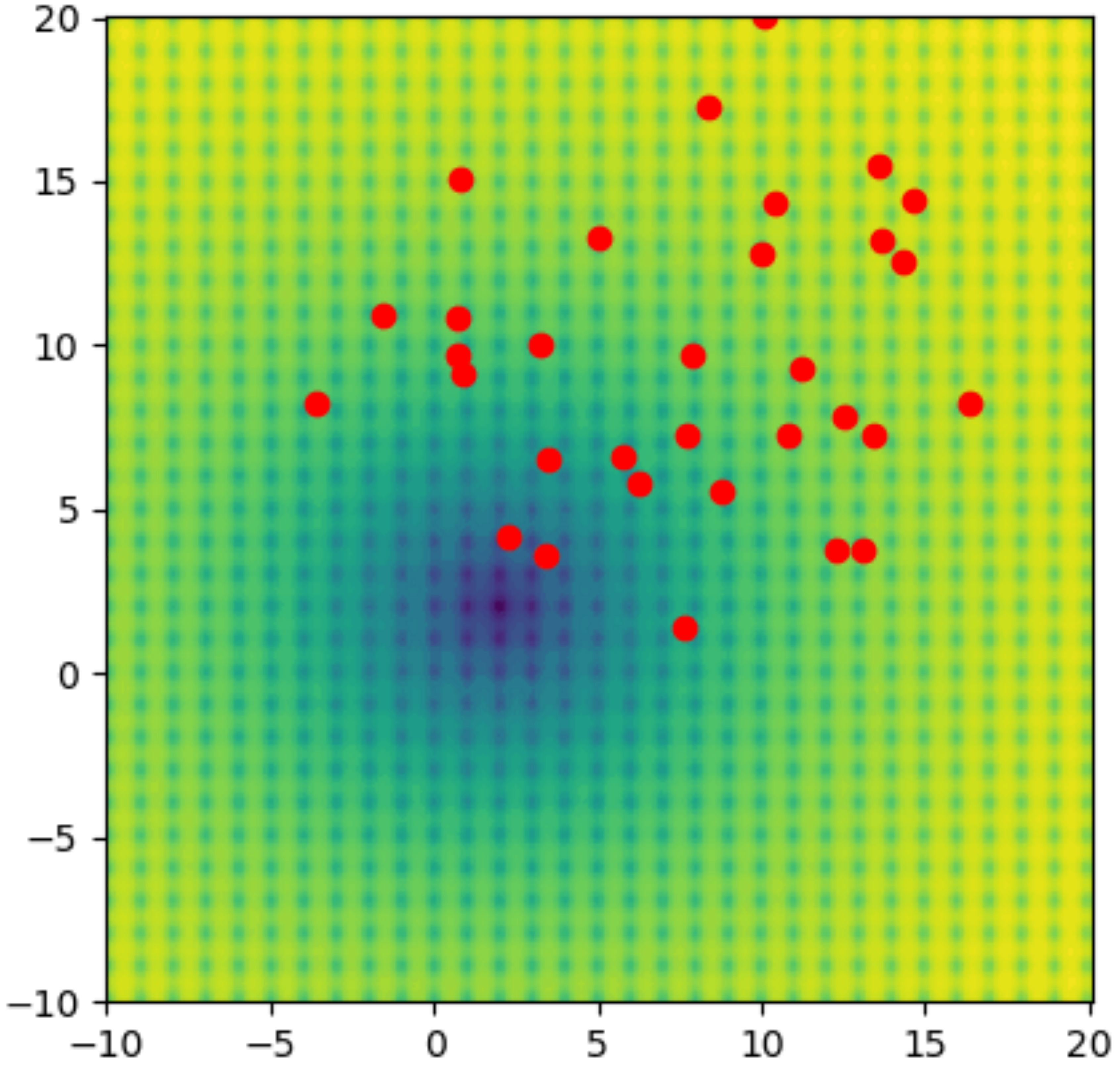
PEDS



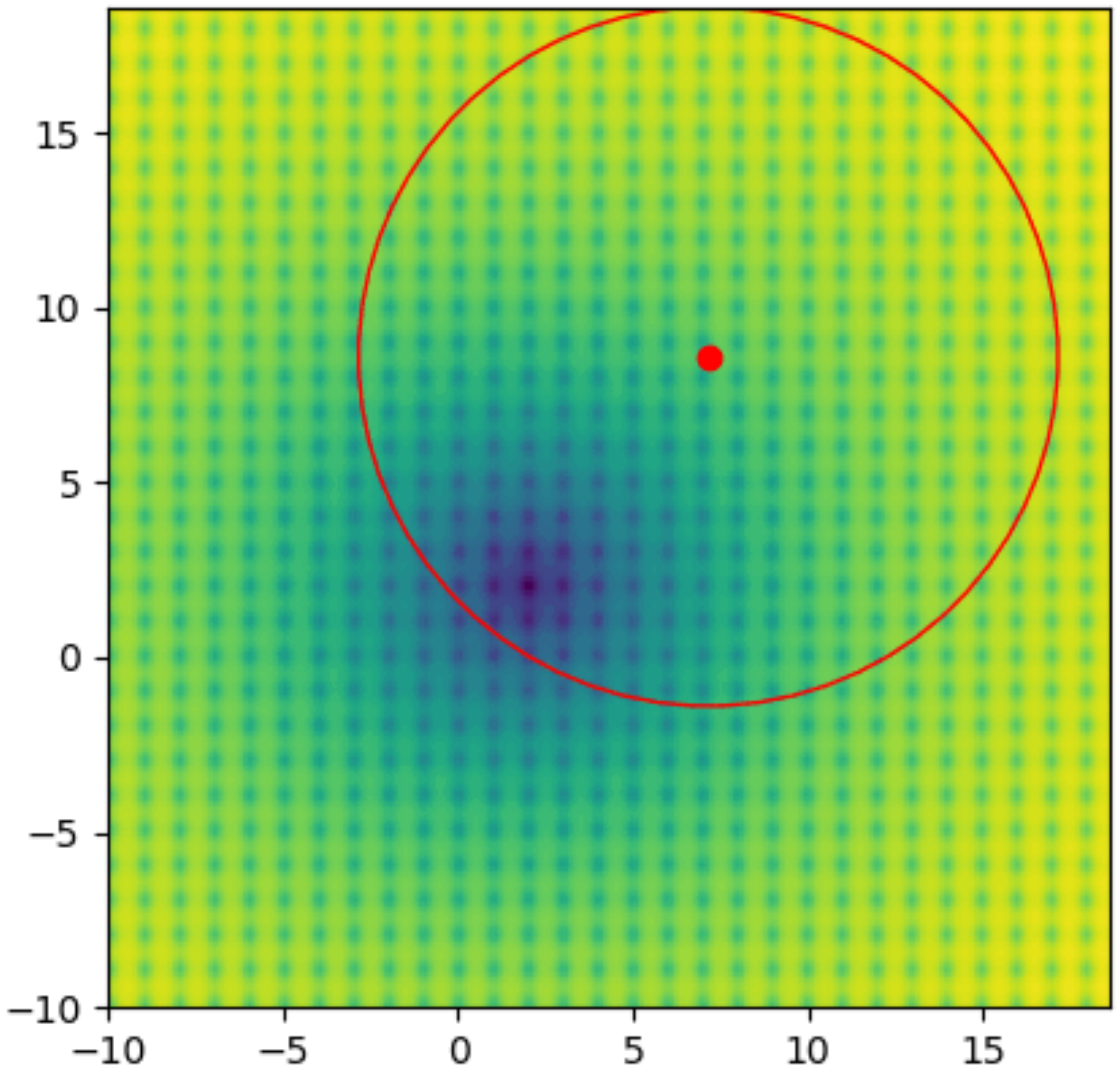
SA-PEDS

Experiments (m=10, N=20)

Only showing first two coordinates, instead of all 10 coordinates.



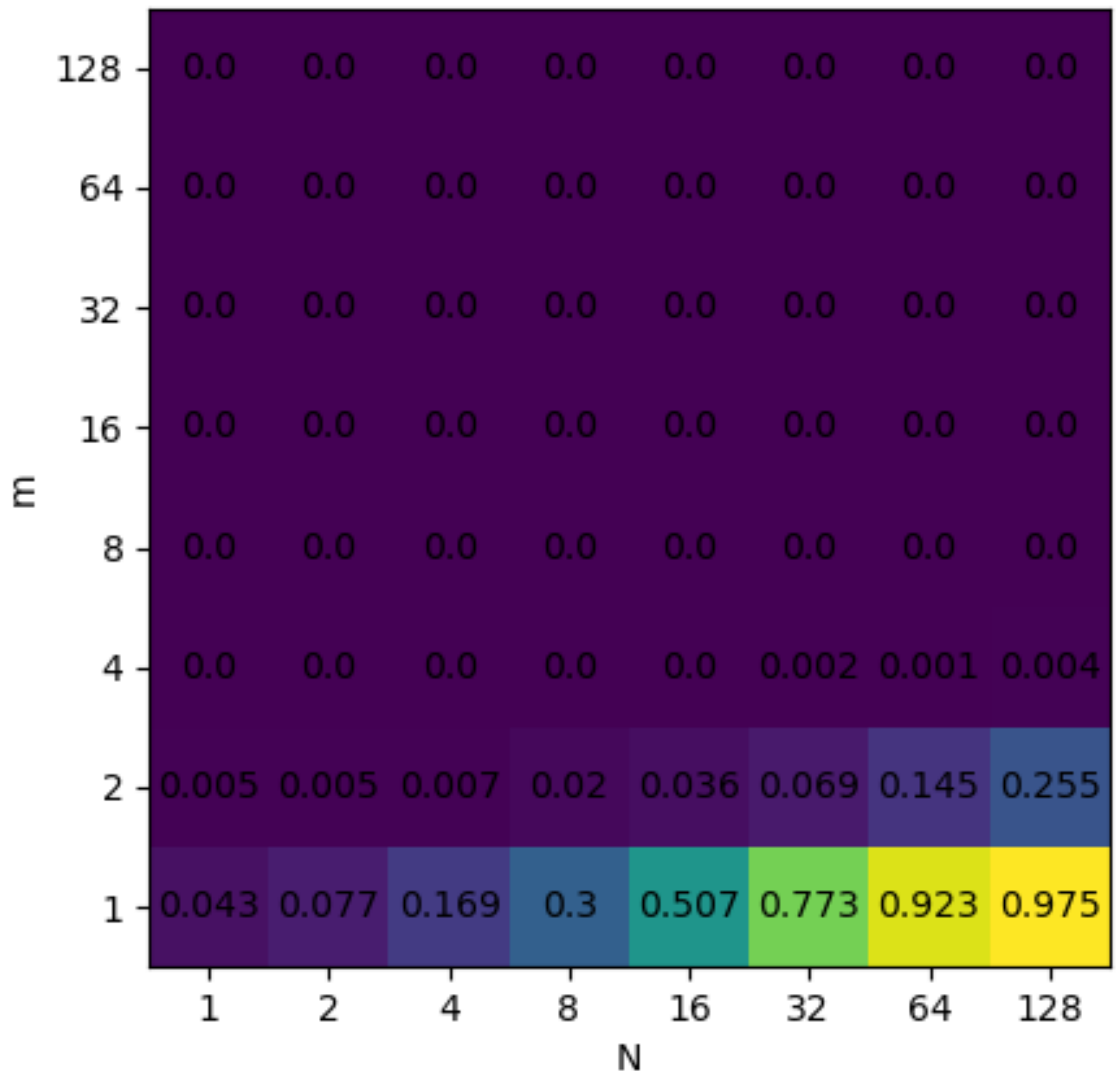
PEDS



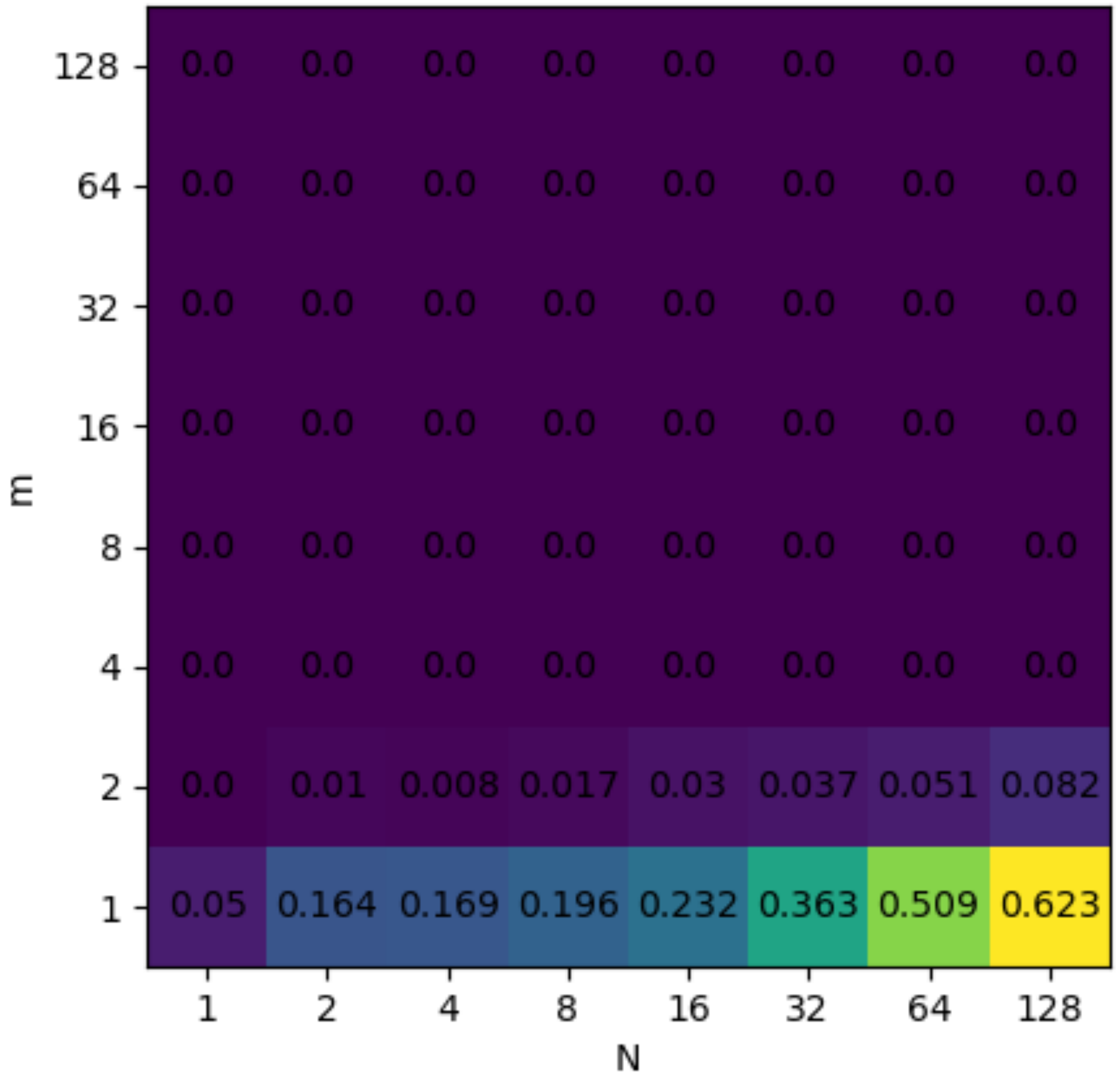
SA-PEDS

Experiments Results

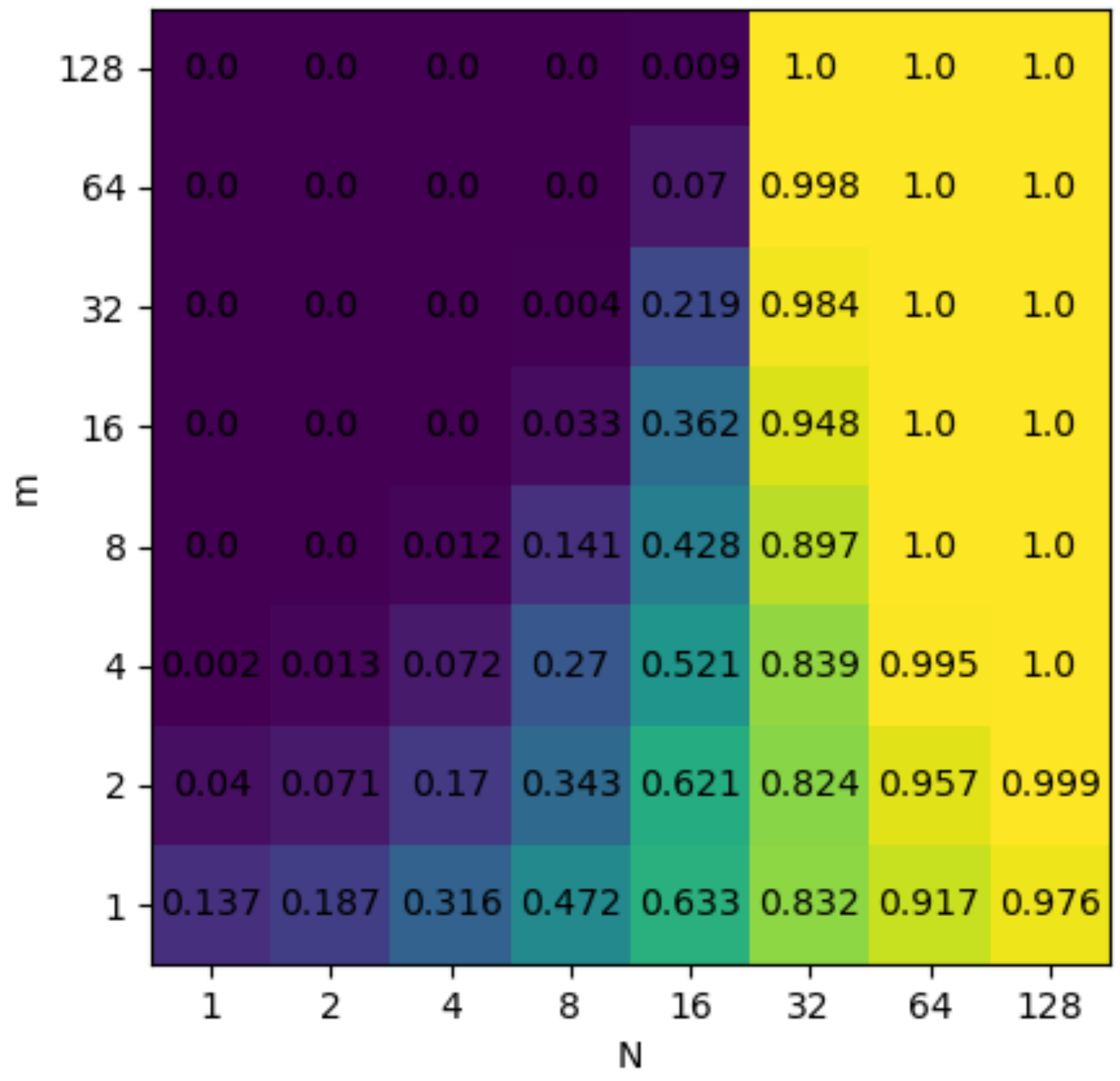
The success rate on increasing N (x-axis) and m (y-axis)



Restart



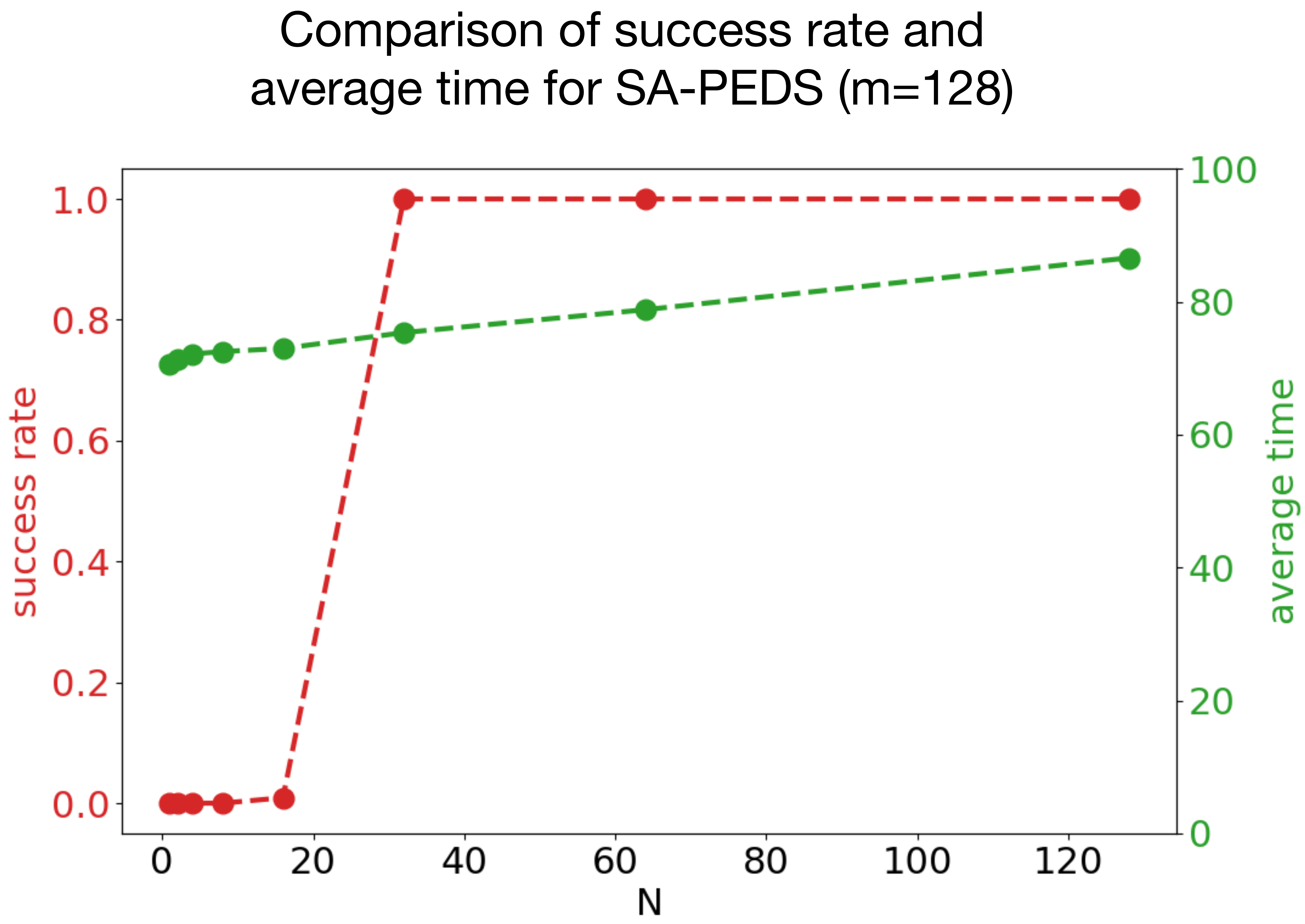
PEDS



SA-PEDS

Experiments Results

How expensive is SA-PEDS?



Discussions

- Inspired by PEDS, we proposed SA-PEDS, which achieves successful convergence behavior on the Ackley function.
- SA-PEDS is for a particular case of PEDS. It’s not a strict generalization.
- If the signal is in high-frequency (e.g. Rosenbrock function), PEDS and SA-PEDS don’t work (preliminary results).
- PEDS and SA-PEDS are sensitive on the value of α (decreasing rate of variance/the attraction force).
- Study this algorithm using particle theory and send N to infinity.

Acknowledgement

This research would not be possible without the support of

- Courant Institute of Mathematical Science
    - For organizing the events
- Simons Center for Computational Physical Chemistry
    - For funding
- NYU High Performance Computing Greene
    - For computation resources

Reference

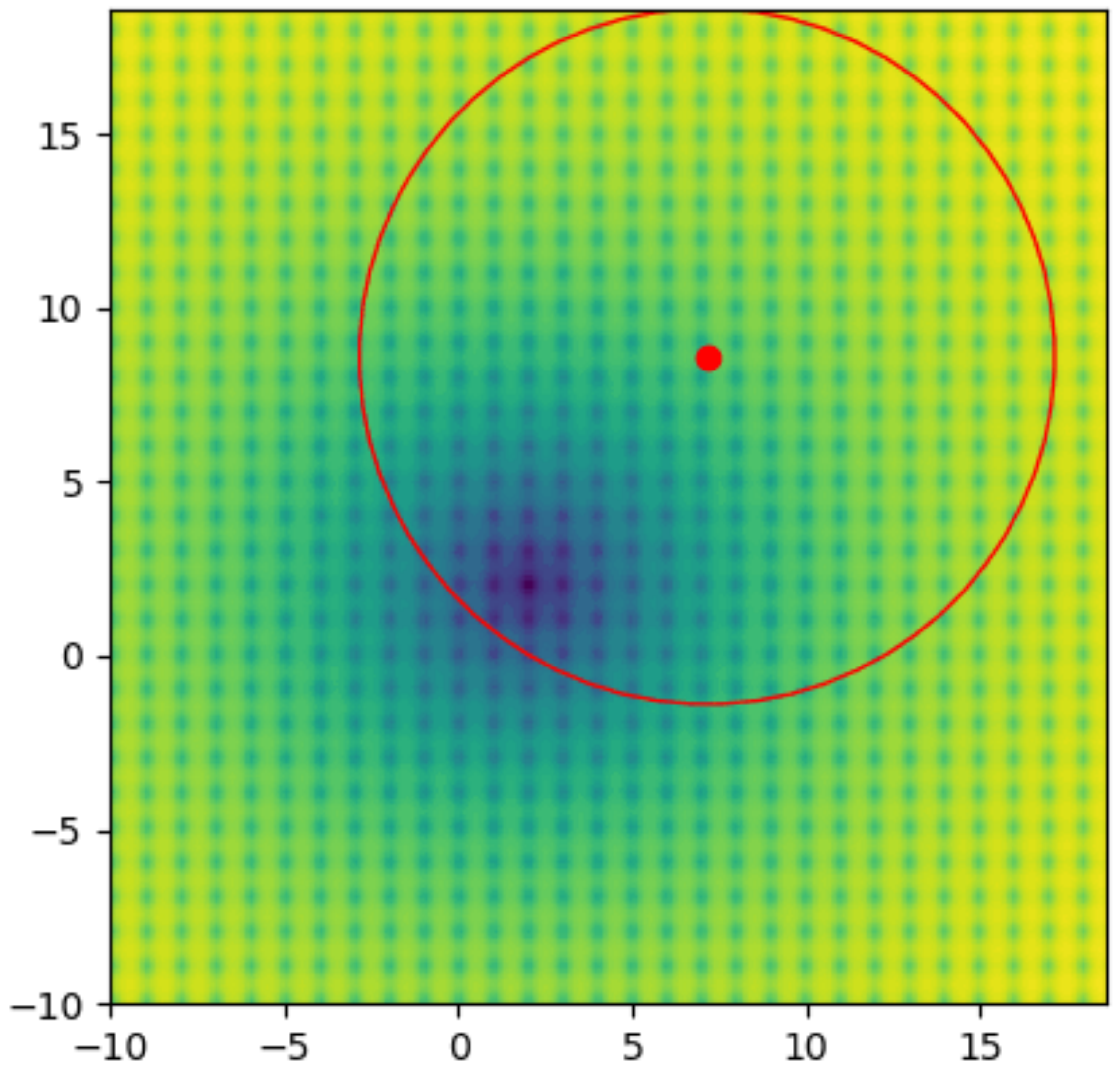
- Caravelli, Francesco, Forrest C. Sheldon, and Fabio L. Traversa. "Global minimization via classical tunneling assisted by collective force field formation." *Science Advances* 7.52 (2021): eabh1542.
- Caravelli, Francesco, et al. "Projective embedding of dynamical systems: Uniform mean field equations." *Physica D: Nonlinear Phenomena* 450 (2023): 133747.
- Duchi, John C., Peter L. Bartlett, and Martin J. Wainwright. "Randomized smoothing for stochastic optimization." *SIAM Journal on Optimization* 22.2 (2012): 674-701.

Appendix A

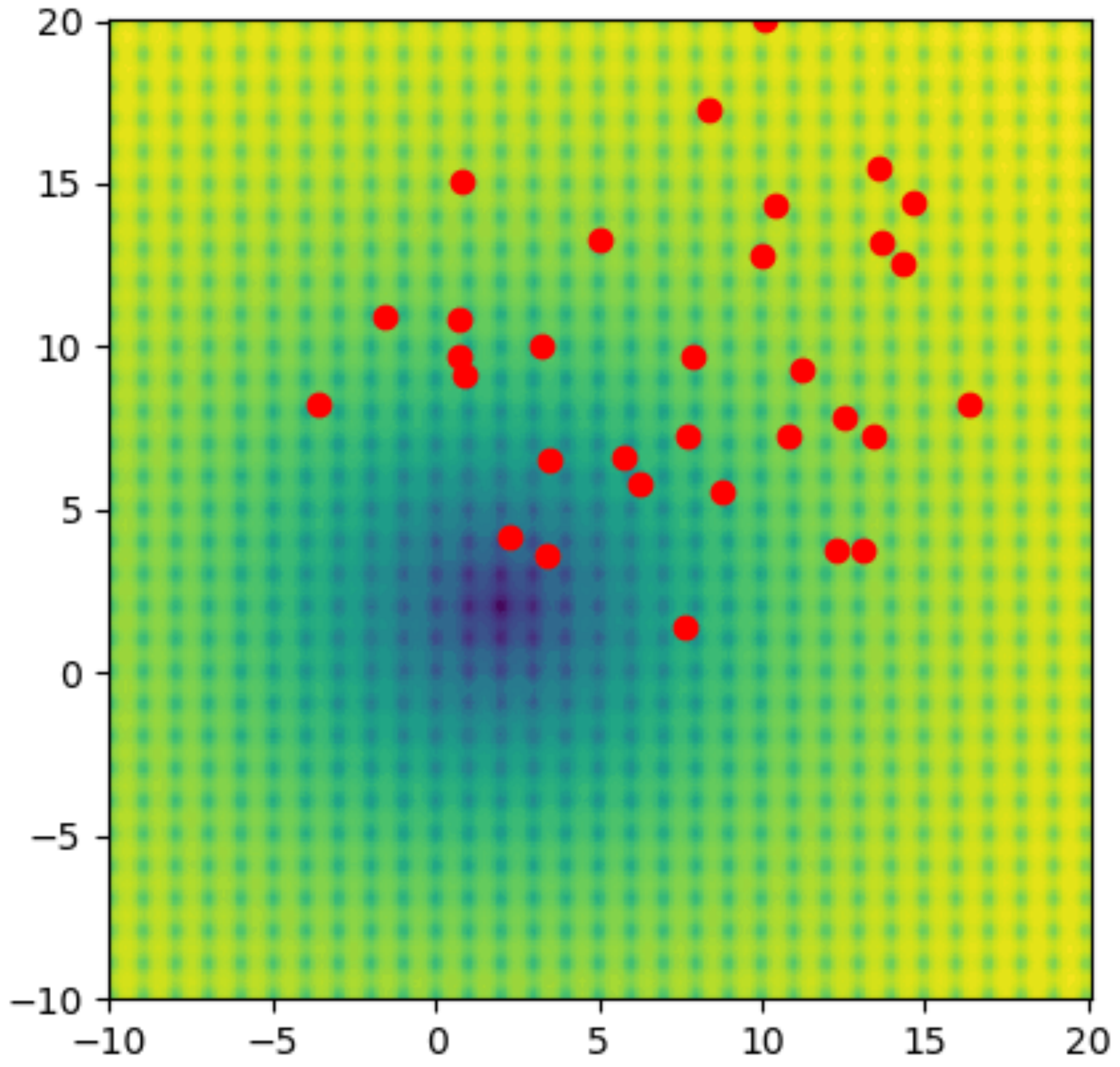
The choice of optimizer

- For SA-PEDS, the variance of gradient will cause large wiggling effect for Vanilla Gradient Descent. Using Adam solves this problem.

- For PEDS with large m, choosing small α and using Adam improves the result.



SA-PEDS with VGD



PEDS with Adam

Appendix B

Accelerating SA-PEDS by importance sampling

- By importance sampling, the expectation of gradient can be evaluated as:
- $\mathbb{E}\nabla F(R) = \begin{pmatrix} \nabla F(R_1) & \nabla F(R_2) & \dots & \nabla F(R_K) \end{pmatrix}^T \begin{pmatrix} \mathcal{N}(R_1;\theta,\sigma) & \mathcal{N}(R_2;\theta,\sigma) & \dots & \mathcal{N}(R_K;\theta,\sigma) \end{pmatrix}$
- After picking a set $\mathfrak{S}$ of points, we can calculate the probability-weighted sum for the expectation. When θ changes a little bit, we can just still get a fairly good approximation by shifting the probability-weight matrix and adding a few new points to $\mathfrak{S}$.
- It's like sliding window / convolution.

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \rightarrow \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & \cdots & & x_{N,m} \end{pmatrix} \; Y_m$$

Appendix C

Some remarks for PEDS

- $Y_j^{t+1} - Y_j^t = -\gamma(\Omega\, \Phi(\nabla F; Y_1^t, Y_2^t, \ldots, Y_m^t) + \alpha(I - \Omega)Y_j^t), j = 1,\ldots,m$

- The original problem in m dimension is embedded into an Nm dimensional space.

- The gradient is projected onto the column space of Ω and the second term, called the **decay function**, ensures that Y_i will also be on the column space of Ω in the long run.

- It is proved that this keeps local minimum and saddle points and it transforms local maximum to be saddle points (Caravelli et al. 2023).

$$\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix} \rightarrow \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ x_{N,1} & \cdots & & x_{N,m} \end{pmatrix} \; R_N \quad Y_m$$

Appendix C

One particular case for PEDS

- $Y_j^{t+1} - Y_j^t = -\gamma \left(\Omega\, \Phi(\nabla F; Y_1^t, Y_2^t, \ldots, Y_m^t) + \alpha(I - \Omega) Y_j^t \right)$
- where $\Phi(\nabla F; Y_1, Y_2, \ldots Y_m)_i = \nabla F \left((m_{i,1}, m_{i,2}, \ldots m_{i,m})^T \right) = \nabla F(R_i)$,
  - $\Omega = \Omega_1 = \dfrac{1}{N} \begin{pmatrix} 1 \cdots 1 \\ \vdots \; \ddots \; \vdots \\ 1 \cdots 1 \end{pmatrix}.$