Visualizing and Understanding the Performance of Randomized Optimization Methods

Rebecca Alexandra Lopez†⋆ (rlopez01@uw.edu), Matt Menickelly⋆, Stefan M. Wild⋆

⋆ Mathematics and Computer Science Division, Argonne National Laboratory

† Department of Applied Mathematics, University of Washington

Motivation

Our goal is to develop a better understanding of optimization algorithms that contain some form of randomization in their inner workings. We develop different ways to visualize the mechanics of an optimization algorithms as it navigates through a decision space, which can be especially challenging as the dimension of the decision space grows.

Background

Given a starting point $\boldsymbol{x}_0 \in \mathbb{R}^n$, most iterative optimization algorithms can be expressed in the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \boldsymbol{d}_k, \qquad k = 0, 1, 2, \ldots, \tag{1}$$

with $\boldsymbol{d}_k \in \mathbb{R}^n$ serving as a scaled direction and α_k being a (nonnegative) step size along $\boldsymbol{d}_k$. By *randomized*, we refer to the fact that some component of the algorithm's mechanics depends on one or more random variables.

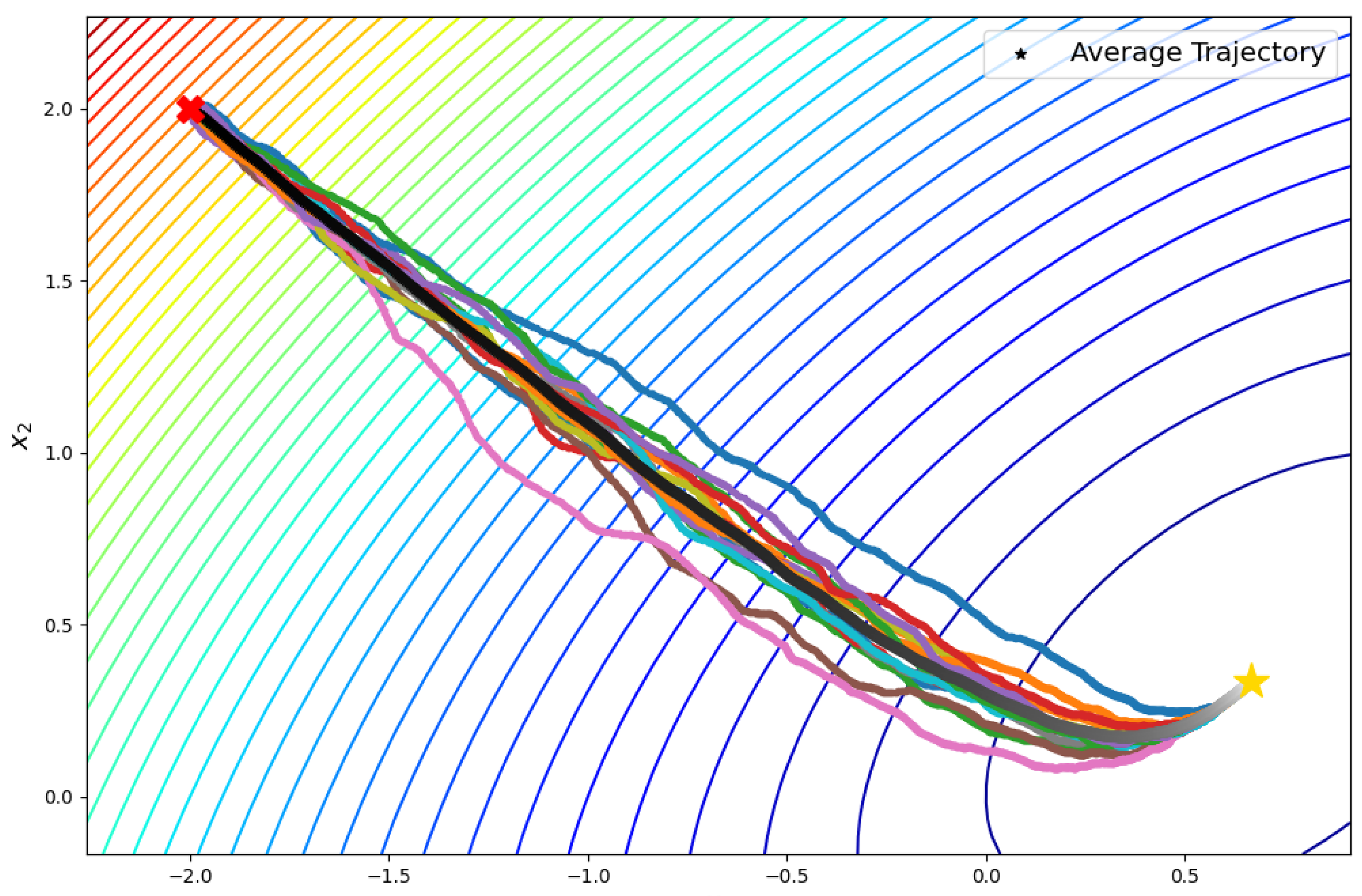


Figure 1: This plot displays a two dimensional convex Nesterov quadratic function using RCD and the Adam optimization method for 15 seeds and 2500 iterations.

We utilize two methods:

1. Let $\boldsymbol{\xi}_k \sim \mathcal{N}(\boldsymbol{0}, \mathbf{I}_n)$ be a n-dimensional vector of standard Gaussians and then set

$$\boldsymbol{d}_k(\boldsymbol{\xi}_k) = \left(\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k)^\top \boldsymbol{\xi}_k\right) \boldsymbol{\xi}_k. \tag{2}$$

2. Randomized coordinate descent (RCD) where one selects an index $\xi_k \sim \mathcal{U}[1, \ldots, n]$ and moves only along this random coordinate:

$$\boldsymbol{d}_k(\xi_k) = \left(\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k)^\top \boldsymbol{e}_{\xi_k}\right) \boldsymbol{e}_{\xi_k} = \frac{\partial}{\partial x_{\xi_k}} f(\boldsymbol{x}_k) \boldsymbol{e}_{\xi_k}, \tag{3}$$

where $\boldsymbol{e}_i \in \mathbb{R}^n$ denotes the ith column of the n-dimensional identity matrix $\mathbf{I}_n$.

The step size is typically chosen ahead of time. One common choice is to set $\alpha_k = \frac{M}{k+1}$ for all k for some $M > 0$. Another choice is to select step sizes adaptively based on estimates of second moments, as in the Adam method [1].

[1] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[2] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87 of *Applied Optimization*. Springer, 2004. doi: 10.1007/978-1-4419-8853-9.

[3] S. U. Stich, C. L. Müller, and B. Gärtner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23 (2):1284–1309, 2013. doi: 10.1137/110853613.

Test Functions

Chained alternative Rosenbrock function:

$$f(\boldsymbol{x}) = \sum_{i=1}^{n-1} \left(100\left(x_i^2 - x_{i+1}\right)^2 + (x_i - 1)^2\right), \tag{4}$$

with $f(\boldsymbol{x}_*) = 0$ at the unique point $\boldsymbol{x} = \mathbf{1}$.

Convex Nesterov [2] quadratic function:

$$f(\boldsymbol{x}) = \frac{1}{2}x_1^2 + \frac{1}{2}x_n^2 - x_1 + \frac{1}{2}\sum_{i=1}^{n-1}(x_{i+1} - x_i)^2, \tag{5}$$

which has $f(\boldsymbol{x}_*) = 0$ at the unique point defined by $x_i = 1 - \frac{i}{n+1}$ for $i = 1, \ldots, n$.

Convex funnel function [3]:

$$f(\boldsymbol{x}) = \log\left(1 + 10\|\boldsymbol{x} - \mathbf{1}\|_2^2\right), \tag{6}$$

which has $f(\boldsymbol{x}_*) = 0$ at the unique point $\boldsymbol{x} = \mathbf{1}$.

Two-Dimensional Functions

Consider the comparison between the step size of $\alpha_k = \frac{m}{k+1}$ and the Adam method in Figure 2. Though both eventually get us to a neighborhood of the minima, the Adam method does it with less variability, and fewer iterations.

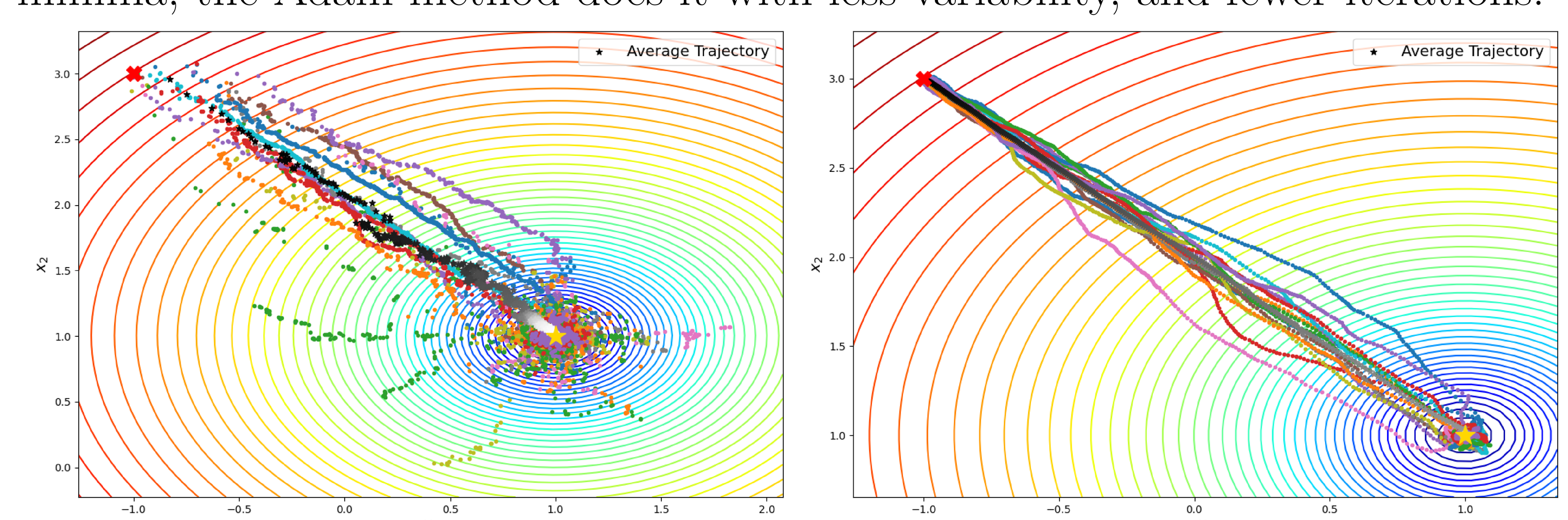


Figure 2: This figure displays the trajectories for 15 seeds and iterations for two optimization methods: (Left) Standard method for selection of α_k, (Right) Adam Optimization Method.

To understand the variability, we calculate standard deviation for each of our $\boldsymbol{x}$ values, essentially creating a two-dimensional confidence interval around each point. However, the standard deviation approach assumes that all components of $\mathbf{x}$ are independent. And so, we look at the covariance confidence ellipses.

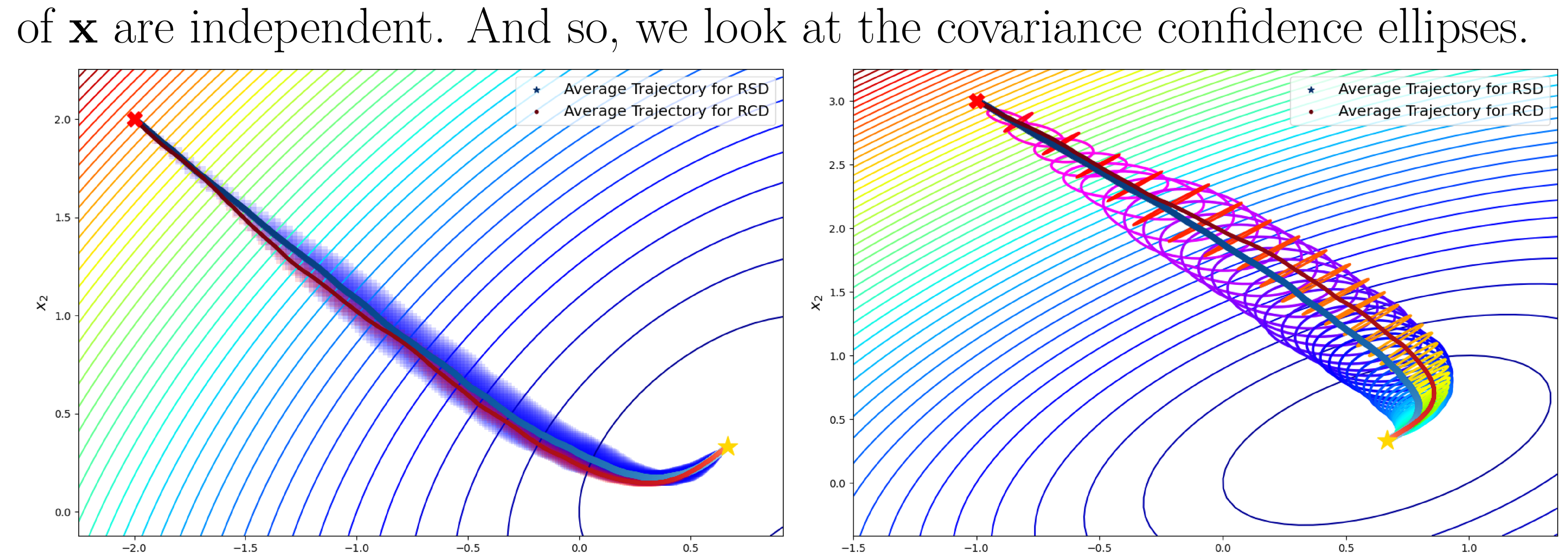


Figure 3: This figure displays the mean trajectories over 15 seeds and iterations for both randomized direction methods: (Left) Standard deviation confidence intervals, (Right) Covariance confidence ellipses.

We would like to acknowledge and thank the Argonne National Laboratory & The National GEM Consortium.

N-Dimensional Functions

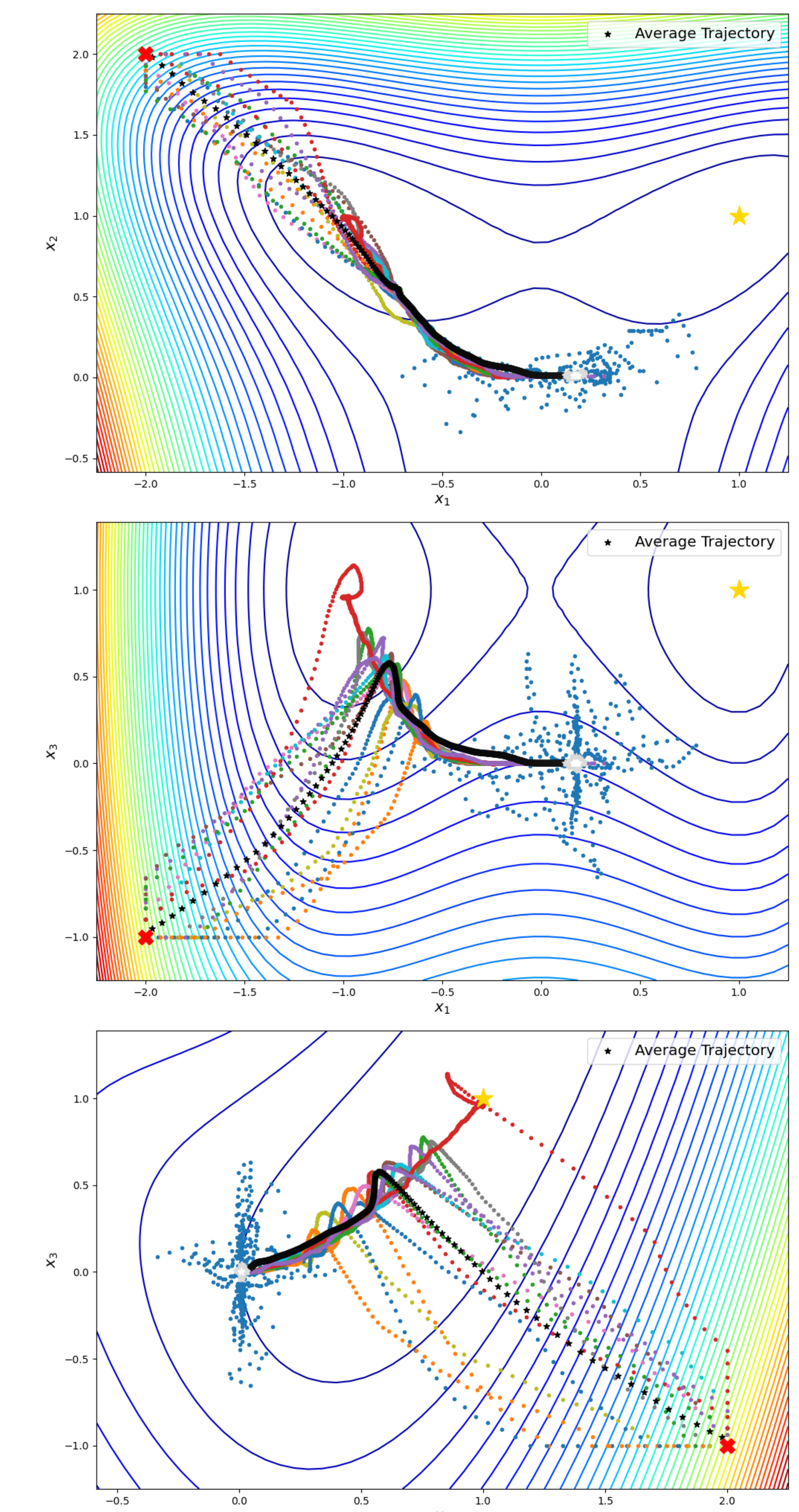
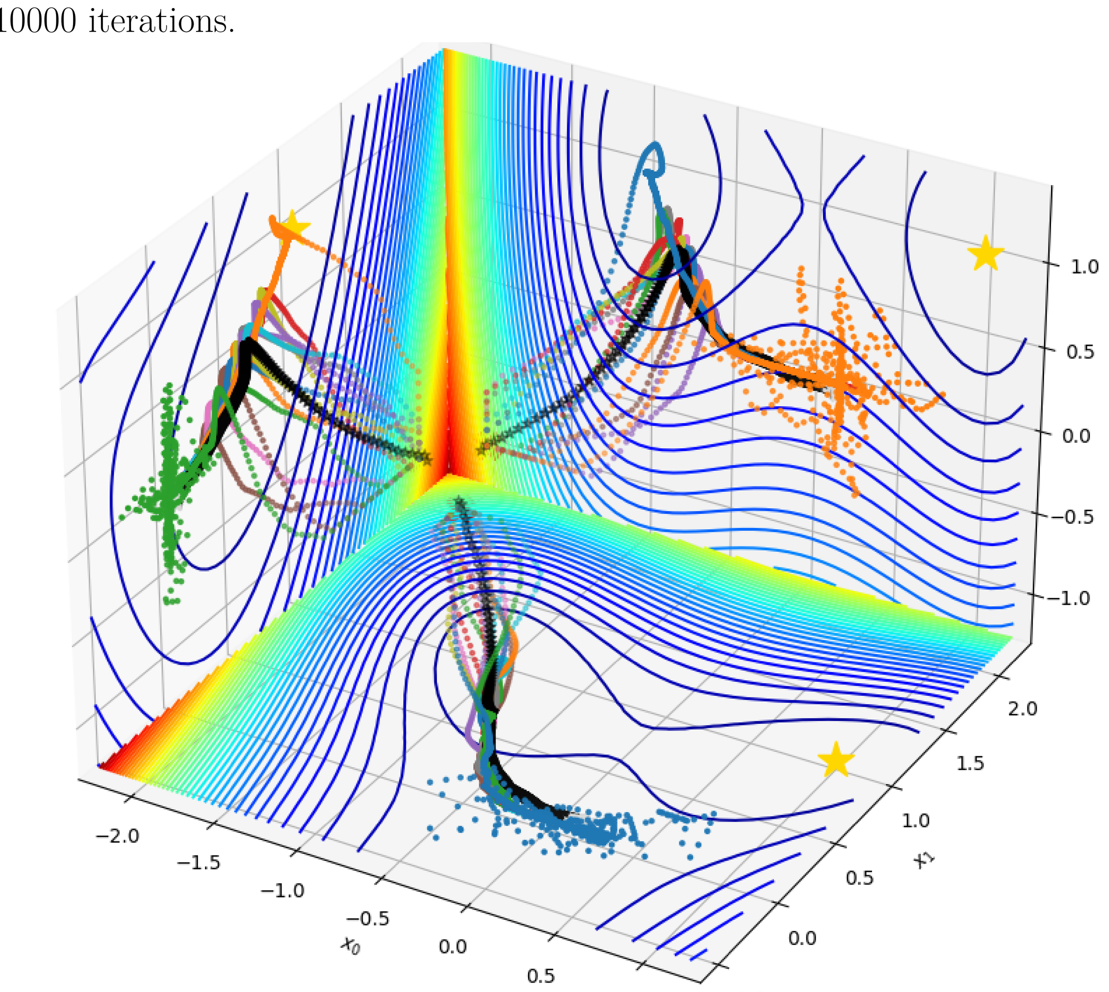


Figure 4: This plot displays all 15 seeds and their mean trajectory along two dimensional slices for the chained alternative Rosenbrock function using RCD and the Adam optimization method with 10000 iterations.

With functions of higher dimension, we reproduce the plots by taking two dimensional slices of each possible combination for $x_i, x_j \in \boldsymbol{x}$ of our trajectories such that $i, j \in 1, 2, \ldots, n$. We produce contours that project on to the two dimensional plane of x_i, x_j by substituting a fixed value equal to the minima for such elements in $\boldsymbol{x}$ that are not x_i, x_j. This allows us to visually characterize the role each element of $\boldsymbol{x}$ plays in the overall trajectory.

Below in Figure 5, we project each slice onto their respective planes to understand and view their relationships.

Figure 5: This plot displays all 15 seeds and their mean trajectory along the two dimensional slices projected along their respectiv planes for the chained alternative Rosenbrock function using RCD and the Adam optimization method with 10000 iterations.



Future Work

- Conduct visualizations on additional nonconvex functions with higher dimensionality
- Apply principal component analysis and produce visualizations for the understanding of such principal components