Unconstrained Global Optimization Using the Adaptive Moment Estimation (ADAM) Algorithm

Ross B. Alexander
Department of Aeronautics and Astronautics
Stanford University
Stanford, CA 94305
rbalexan@stanford.edu

I. Unconstrained Optimization

We performed unconstrained optimization on three optimization benchmark functions – the Rosenbrock function (f_R), the Himmelblau function (f_H), and the Powell function (f_P) [1]–[3] – using the adaptive moment estimation (ADAM) optimization algorithm [4].

A. Benchmark Functions

The Rosenbrock function $f_R : \mathbb{R}^2 \to \mathbb{R}$, is a multimodal, non-convex, two-dimensional function with a steep valley. The floor of the valley is very flat and can cause problems with algorithms utilizing the gradient alone. The Himmelblau function $f_H : \mathbb{R}^2 \to \mathbb{R}$ is also a multimodal, non-convex, two-dimensional function, though it has several local minima – each at $f(\mathbf{x}^*) = 0$. The Powell function $f_P : \mathbb{R}^4 \to \mathbb{R}$, is a unimodal, four-dimensional function. The benchmark functions are shown in Eqns. (1)-(3).

$$f_R(\mathbf{x}) = 100(x_2 - x_1^2)^2 + (1 - x_1)^2 \tag{1}$$

$$f_H(\mathbf{x}) = (x_1^2 + x_2 - 11)^2 + (x_1 + x_2^2 - 7)^2 \tag{2}$$

$$f_P(\mathbf{x}) = (x_1 + 10x_2)^2 + 5(x_3 - x_4)^2 + (x_2 - 2x_3)^4 + 10(x_1 - x_4)^4 \tag{3}$$

B. Adaptive Moment Estimation (ADAM) Algorithm

We implemented many algorithms, but ultimately settled on the ADAM optimization algorithm for its robust performance across each of the benchmark functions. The ADAM algorithm uses a cumulative average of both the first moment of the gradient and the second moment of the gradient. Eqns. (4) and (5) describe the (biased) estimation step where we construct moving estimates of the first and second moments of the gradient. Eqns. (6) and (7) describe the bias correction step where we correct the biased estimators. And finally, Eqn. (8) describes the update step for iterate $\mathbf{x}^{(k)}$ to $\mathbf{x}^{(k+1)}$, where α is the learning rate that is adapted by our unbiased estimators and where ϵ is set to a small value to prevent division by zero.

$$\mathbf{v}^{(k+1)} = \gamma_v \mathbf{v}^{(k)} + (1 - \gamma_v)\mathbf{g}^{(k)} \tag{4}$$

$$\mathbf{s}^{(k+1)} = \gamma_s \mathbf{s}^{(k)} + (1 - \gamma_s)(\mathbf{g}^{(k)} \odot \mathbf{g}^{(k)}) \tag{5}$$

$$\hat{\mathbf{v}}^{(k+1)} = \mathbf{v}^{(k+1)} / (1 - \gamma_v^{k+1}) \tag{6}$$

$$\hat{\mathbf{s}}^{(k+1)} = \mathbf{s}^{(k+1)} / (1 - \gamma_s^{k+1}) \tag{7}$$

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \hat{\mathbf{v}}^{(k+1)} / \left(\sqrt{\hat{\mathbf{s}}^{(k+1)}} + \epsilon \right) \tag{8}$$

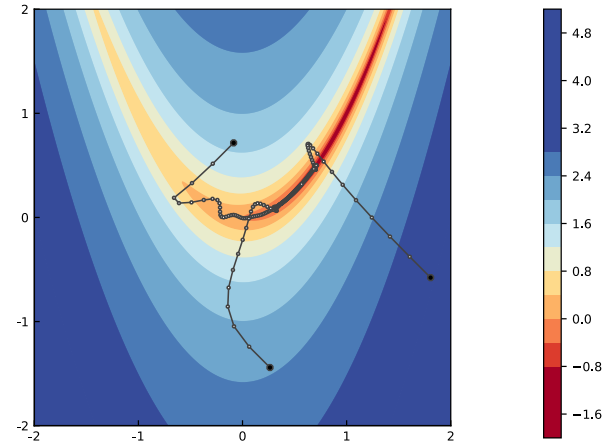

Fig. 1. Three optimization paths on the Rosenbrock function. The ADAM optimizer was used with a learning rate of $\alpha = 0.2$, a first moment estimate decay factor $\gamma_v = 0.7$, and a second moment estimate decay factor $\gamma_s = 0.99$. The color gradations are the base-10 logarithm of the function values.

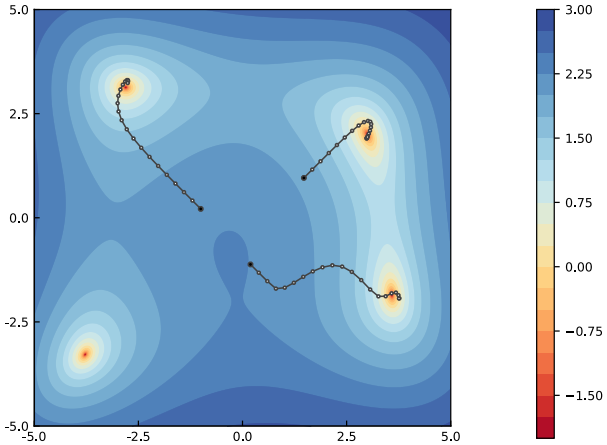

Fig. 2. Three optimization paths on the Himmelblau function. The ADAM optimizer was used with a learning rate of $\alpha = 0.2$, a first moment estimate decay factor $\gamma_v = 0.7$, and a second moment estimate decay factor $\gamma_s = 0.99$. The color gradations are the base-10 logarithm of the function values.

C. Results & Discussion

Figures 1 and 2 show three randomly-initialized optimizations on the Rosenbrock and Himmelblau functions, respectively. We can see that each of the optimizations traverses the local gradient and utilizes the momentum ($\mathbf{v}$) to improve

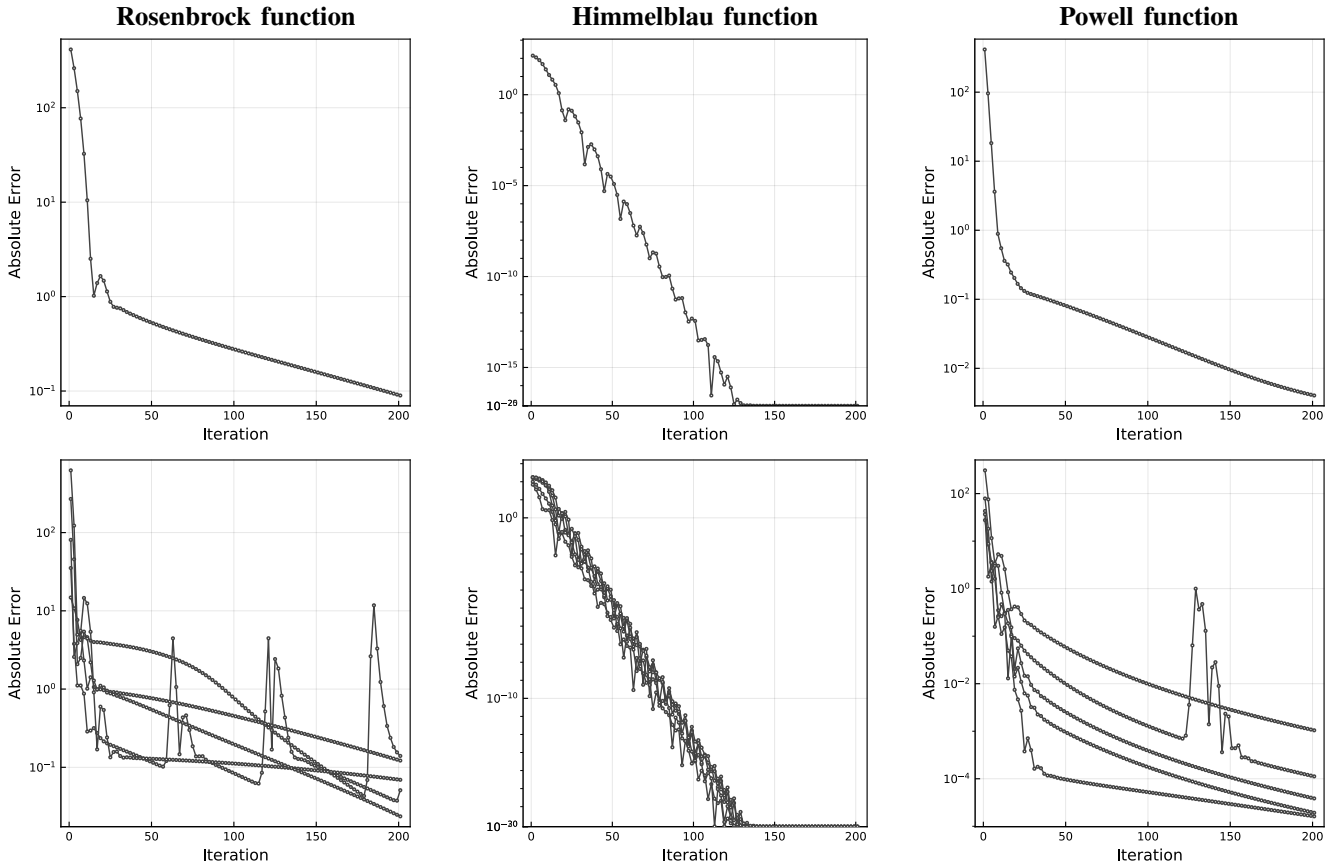

Fig. 3. Convergence of the absolute error in optimizing the Rosenbrock function (*left*), the Himmelblau function (*center*), and the Powell function (*right*). The upper figures depict a single randomly-initialized optimization and the lower figures depict a series of five randomly-initialized optimizations, demonstrating consistency in the convergence to a global minimum.

performance in regions with a shallow gradient. The three optimizations on the Himmelblau function highlight the responsiveness to local conditions as three random initializations converge to local minima in their neighborhood.

We show several absolute error convergence plots for the Rosenbrock, Himmelblau, and Powell functions in Fig. 3. The upper row of figures shows the convergence for a single randomly-initialized optimization, while the lower row of figures shows the convergence for five randomly-initialized optimizations.

Optimizations on the Rosenbrock function are generally slow to converge, as shown in the very shallow rate of convergence, likely due to the flatness of the valley in which the global minimum lies. The Rosenbrock optimizations are also less consistent than any other function optimizations, showing various convergence rates along with some oscillatory behavior for one of the optimizations.

Optimizations on the Himmelblau function show the robustness of the ADAM optimization method with impressive convergence rates and consistency across iterations. We attribute this to the relatively simple and unimodal regions containing the local minima, which make convergence easily attainable.

The four-dimensional Powell function displays similar convergence rates as the two-dimensional Rosenbrock function, despite the increased dimensionality of the search space. While the optimizations seem to have different "burn-in" periods, the optimizations have consistent convergence rates (aside from an unusual peak in one of the optimizations).

Overall, the adaptive moment estimation (ADAM) optimization method was a robust and efficient choice for optimization on these benchmark functions along with two secret benchmark functions in the autograder. It would be interesting to investigate the performance of some newer variants of ADAM with Nesterov momentum (NADAM) [5] and with improved convergence guarantees (AMSGRAD) [6].

REFERENCES

[1] H. H. Rosenbrock, "An Automatic Method for Finding the Greatest or Least Value of a Function," *The Computer Journal*, 1960.

[2] D. M. Himmelblau, *Applied Nonlinear Programming*. McGraw-Hill, 1972.

[3] M. J. D. Powell, "An Iterative Method for Finding Stationary Values of a Function of Several Variables," *The Computer Journal*, 1962.

[4] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[5] T. Dozat, "Incorporating Nesterov Momentum into Adam," *ICLR Workshop*, 2016.

[6] S. J. Reddi, S. Kale, and S. Kumar, "On the convergence of Adam and beyond," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.