Variational Optimization Landscapes

William Saunders

Department of Computer Science
University of Toronto

Jan 26, 2018

The upper bound view

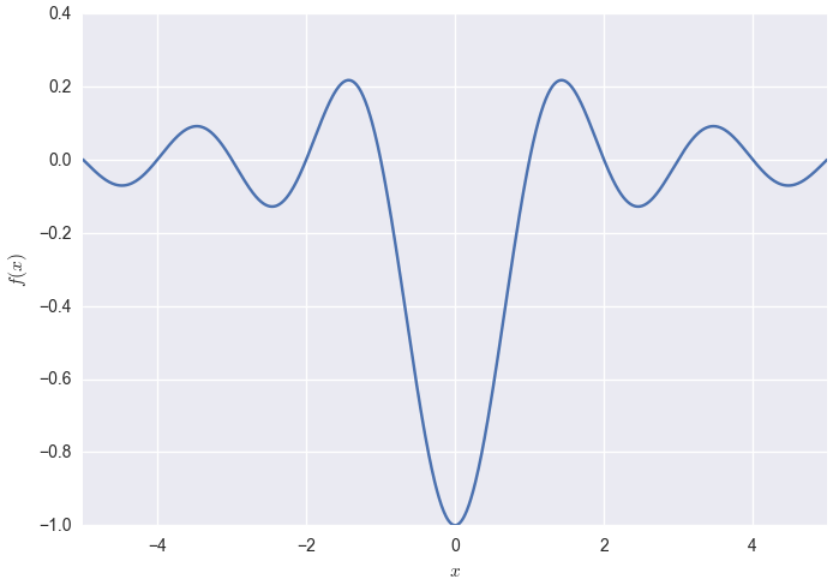
- Instead of directly minimizing the function, we can minimize the upper bound:
  $\min_x f(x) \leq \min_\theta \mathbb{E}_{x \sim p(x|\theta)}[f(x)]$
- often, we take $p(x|\theta) = N(\theta, \Sigma)$
- How does this change the optimization landscape and the solutions we find?

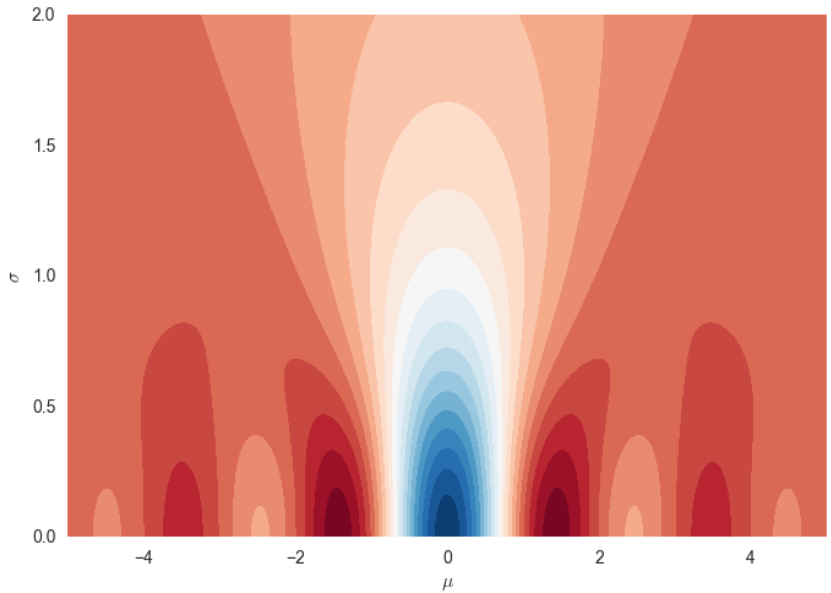
Sources

Evolution Strategies, Variational Optimization and Natural ES
By: Ferenc Huszr
http://www.inference.vc/evolution-strategies-variational-optimisation-and-natural-es-2/

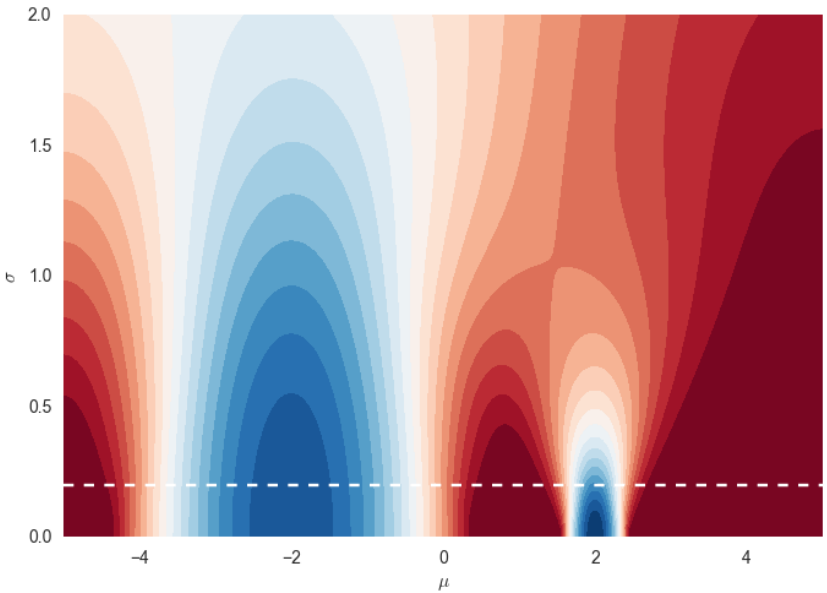
ES is more than just a traditional finite-difference approximator.
Joel Lehman, Jay Chen, Jeff Clune, and Kenneth O. Stanley.

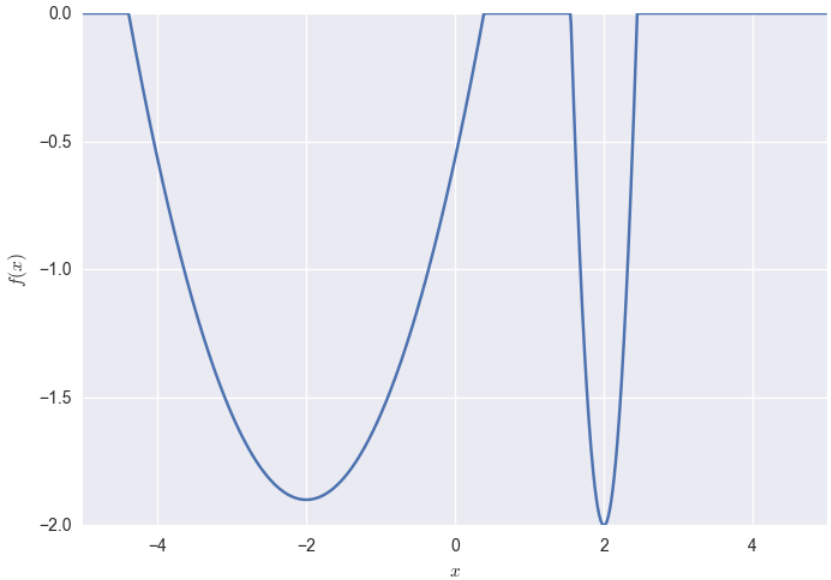
sinc function

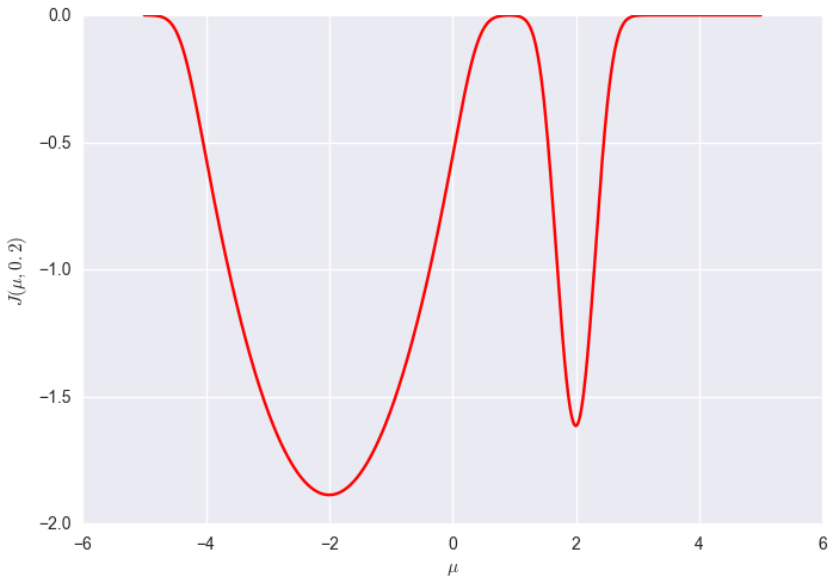




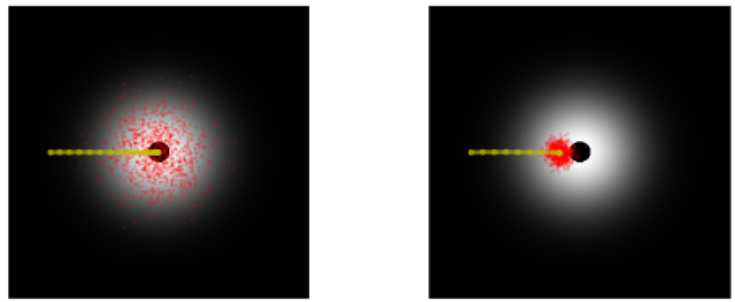
σ too small





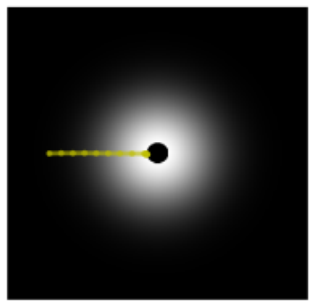


Donut Landscape

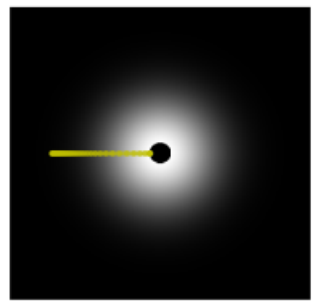


(a) ES with $\sigma = 0.16$

(b) ES with $\sigma = 0.04$



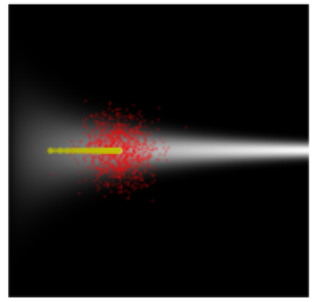
(c) ES with $\sigma = 0.002$



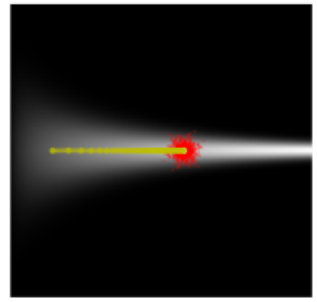
(d) Finite Differences with $\epsilon = 1e-7$

With (a) high variance, ES maximizes expected fitness by moving the distributions mean into a low-fitness area. With (b,c) decreasing variance, ES is drawn closer to the edge of the low-fitness area, qualitatively converging to the behavior of (d) finite-difference gradient descent.

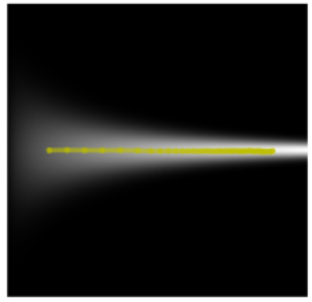
Narrowing Path Landscape



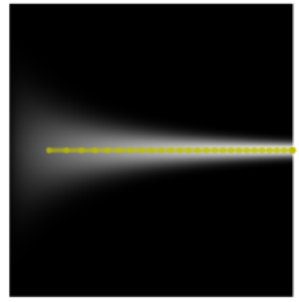
(a) ES with $\sigma = 0.12$



(b) ES with $\sigma = 0.04$



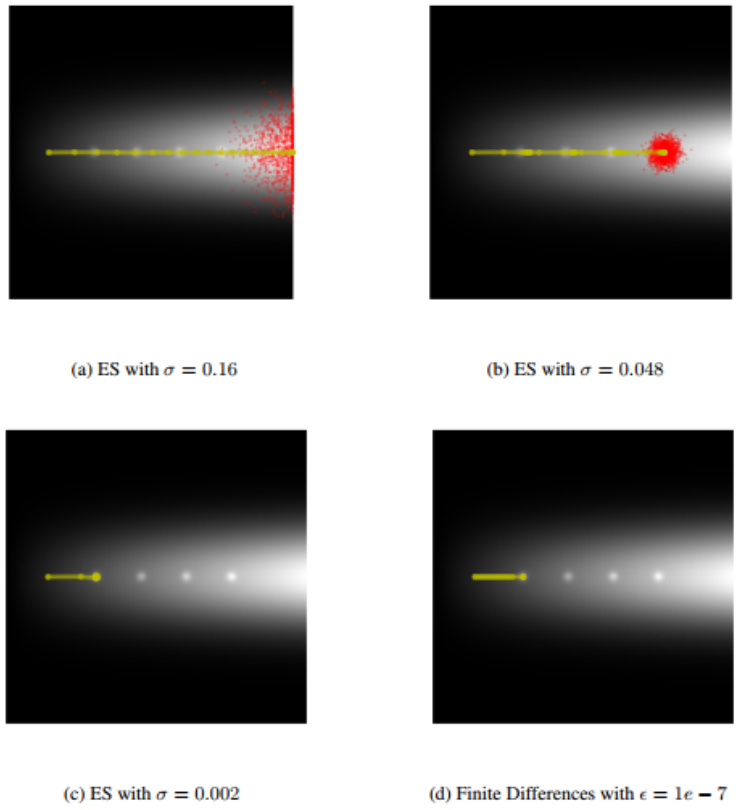
(c) ES with $\sigma = 0.0005$



(d) Finite Differences with $\epsilon = 1e-7$

With (a) high variance, ES maximizes expected fitness by staying on the wider part of the path, meaning it does not traverse the entire path. With (b,c) decreasing variance, ES is able to traverse further along the narrowing path. (d) Finite-difference gradient descent traverses the entire path.

Fleeting Peaks Landscape



(a) ES with $\sigma = 0.16$

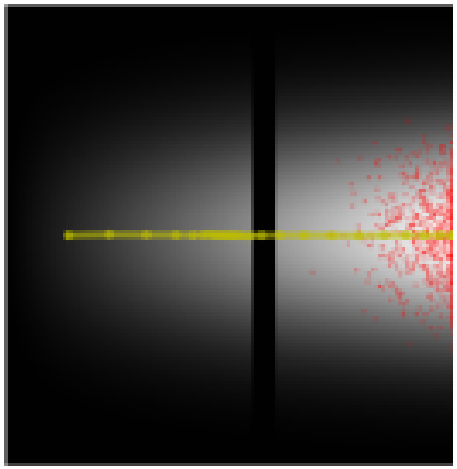
(b) ES with $\sigma = 0.048$

(c) ES with $\sigma = 0.002$

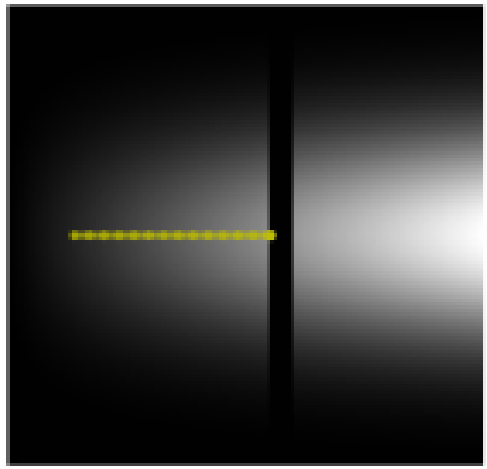
(d) Finite Differences with $\epsilon = 1e - 7$

With (a) high variance, ES can bypass the local optima because its contribution to expected fitness across the distribution is small. With (b) medium variance, ES hops between local optima, and with (c) low variance, ES converges to a local optimum, similarly to (d) finite-difference gradient descent

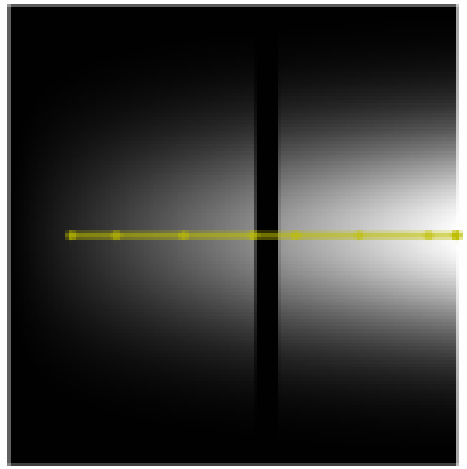
Gradient Gap Landscape



(a) ES with $\sigma = 0.18$



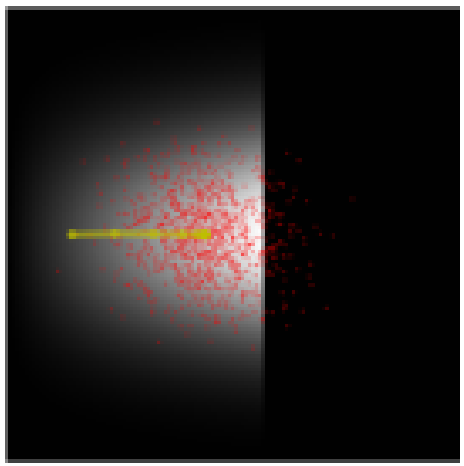
(b) Finite Differences



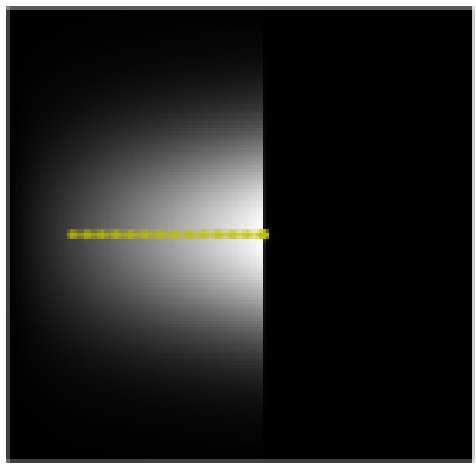
(c) Finite Differences + Momentum

With (a) high variance, ES can bypass the gradient-free gap because its distribution can span the gap; with lower-variance ES or (b) finite differences, search cannot cross the gap. When (c) finite differences is augmented with momentum, it too can jump across the gap.

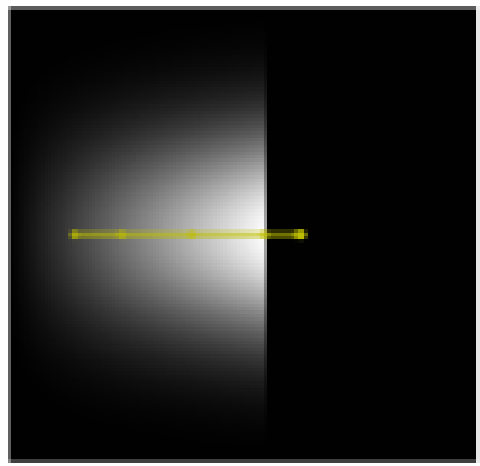
Gradient Cliff Landscape



(d) ES with $\sigma = 0.18$



(e) Finite Differences



(f) Finite Differences + Momentum

In the control Gradient Cliff landscape, (d) ES with high variance remains rooted in the high-fitness area, and the performance of (e) finite differences is unchanged from the Gradient Gap landscape. When (f) finite differences is combined with momentum, it jumps into the fitness chasm.