» Polyak Momentum/Heavy Ball

$x_0 = x0; z_0 = 0; t = 0$

for k in range(num_iters):

$z_{t+1} = \beta z_t + \alpha \nabla f(x_t)$

$x_{t+1} = x_t - z_{t+1}$

$t = t + 1$

* Step z_t at iteration t is weighted sum of past gradients $\alpha \nabla f$:

$$z_{t+1} = \beta^{t-1} \alpha \nabla f(x_0) + \beta^{t-2} \alpha \nabla f(x_1) + \cdots + \alpha \nabla f(x_t)$$

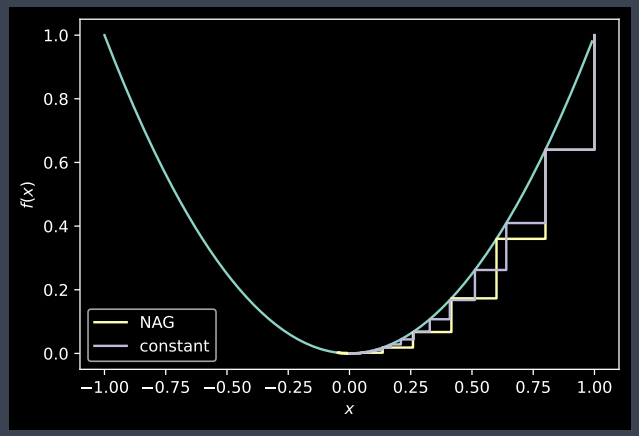
* Need to manually select β and α. A typical choice for β seems to be 0.9, but there is not a universal value and a poor choice can make performance worse than using constant step size
* When $\beta = 0$ recover constant step size strategy
* Since $1 + \beta + \beta^2 + \cdots = 1/(1 - \beta)$, we're effectively using scaling baseline step size $\alpha/(1 - \beta)$ i.e. for $\beta = 0.9$ using 10α

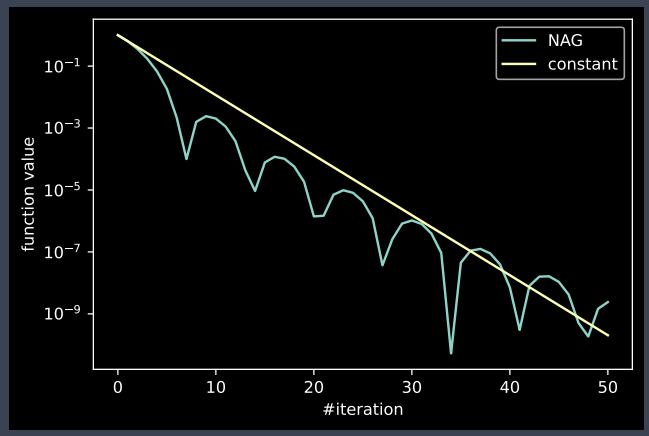
» Nesterov Momentum/Acceleration

* *Nesterov Accelerated Gradient (NAG) Method*, *Nesterov Fast Gradient Method*, *Nesterov Momentum*:
  $z_{t+1} = \beta z_t - \alpha \nabla f(x_t + \beta z_t)$, $x_{t+1} = x_t + z_{t+1}$
* For comparison, heavy-ball momentum uses:
  $z_{t+1} = \beta z_t + \alpha \nabla f(x_t)$, $x_{t+1} = x_t - z_{t+1}$
  $\rightarrow$ substantive difference is that Nesterov changes $\nabla f(x_t)$ to $\nabla f(x_t + \beta z_t)$.
* If $\alpha \approx 0$ then $z_{t+1} \approx \beta z_t$ and $x_{t+1} = x_t + z_{t+1} \approx x_t + \beta z_t$, so $\nabla f(x_t + \beta z_t)$ is "looking ahead" or predicting where x_t is roughly expected to be at next step.
* Need to manually select β and α. A typical choice for β is $(t-1)/(t+2)$ i.e. β changes at each time step with $\beta \to 1$ as $t \to \infty$. Sometimes cap β at e.g. 0.9 or 0.95
* Nesterov acceleration comes with some theoretical guarantees on convergence rate $\rightarrow$ will come back to how useful such guarantees are in practice later

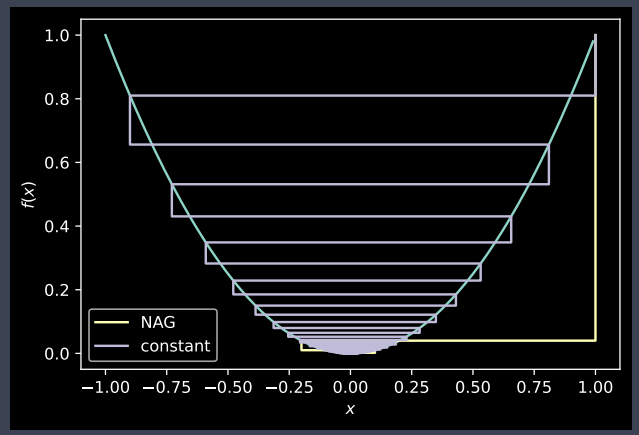
» Examples

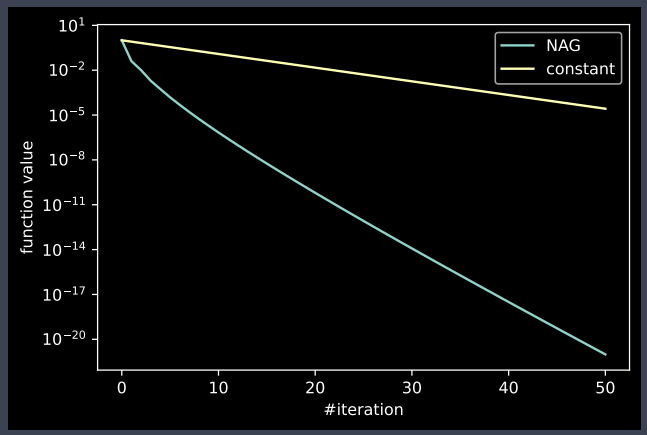
* Example: $f = x^2$, starting point $x = 1$
* Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.1$, constant $\alpha = 0.1$





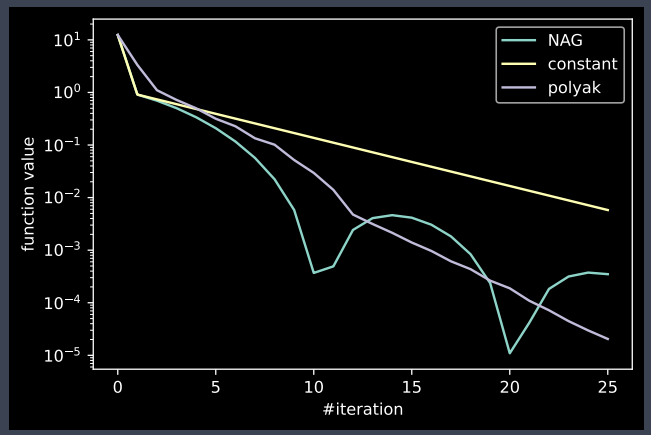
* Again, "bumpy" behaviour of function value vs time when using momentum. Nesterov convergence rate about the same as for constant step size, but can increase step size. Nesterov $\alpha = 0.6$, constant $\alpha = 0.95$:

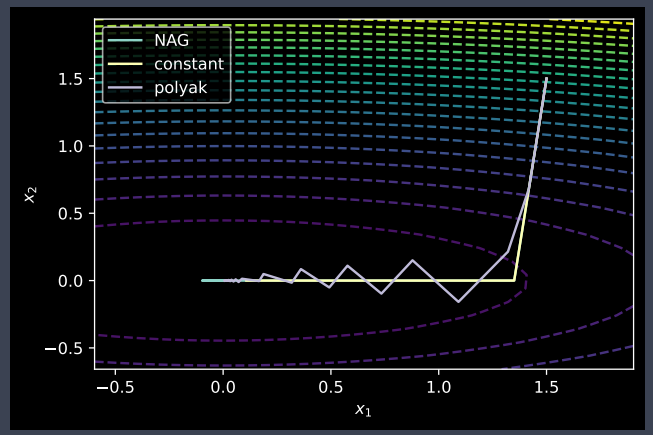




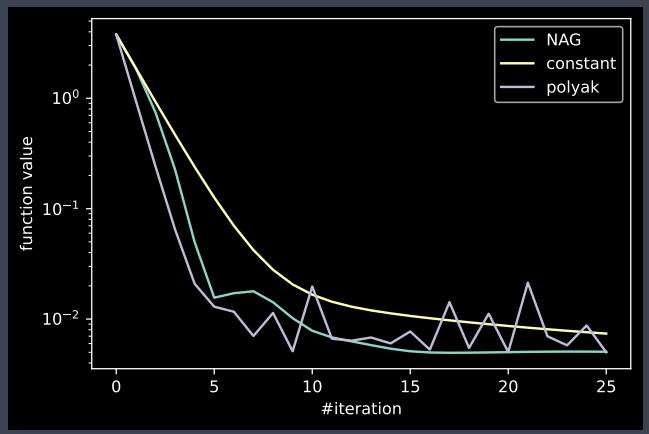
» Examples

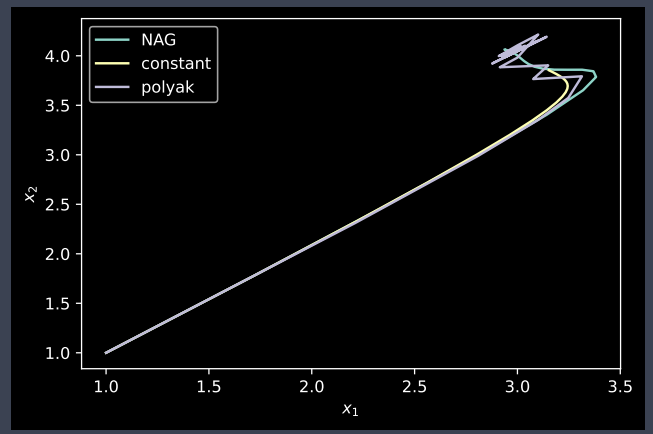
* Quadratic: Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.1$, constant $\alpha = 0.1$





* Quadratic Loss: Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.5$, constant $\alpha = 0.5$

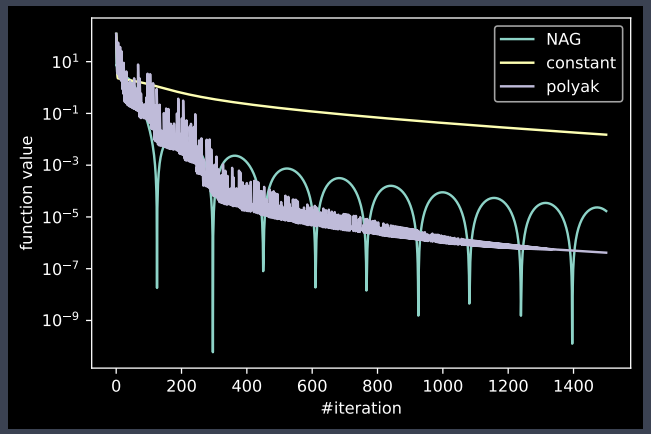


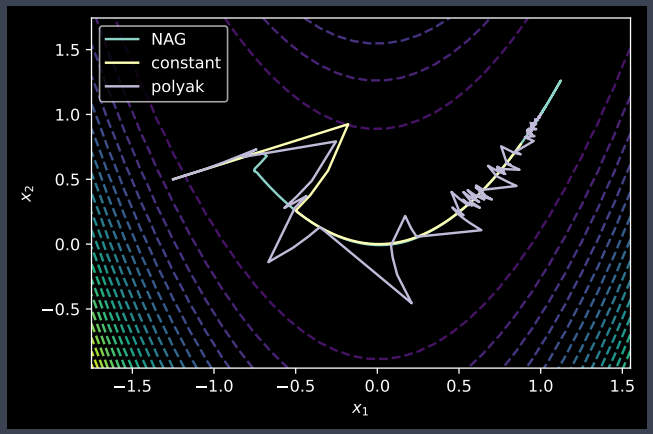


* Nesterov convergence rate similar to Polyak

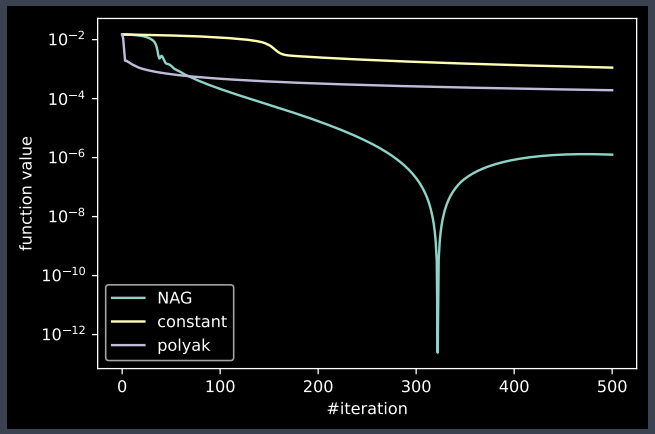
» Examples

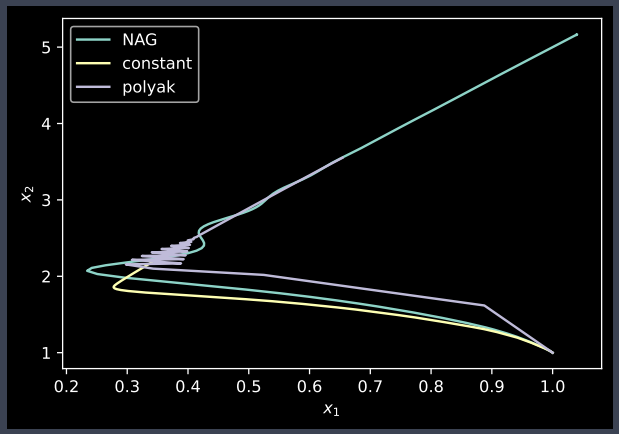
* Rosenbrock function: Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.001$, constant $\alpha = 0.002$





* Toy neural net loss: Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.5$, constant $\alpha = 0.75$

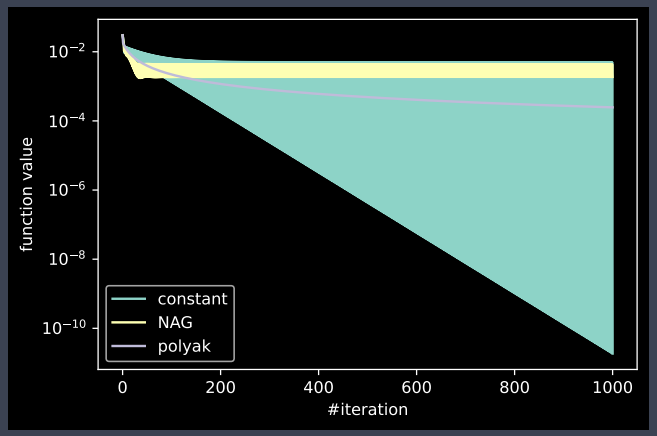


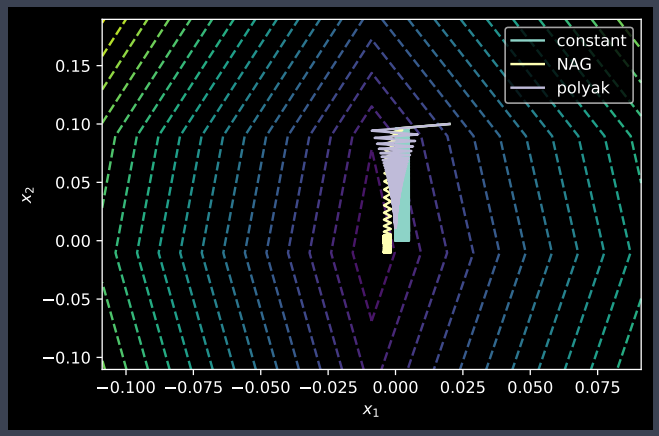


* For Rosenbrock function Nesterov convergence rate similar to Polyak, initially slower for toy neural net then faster

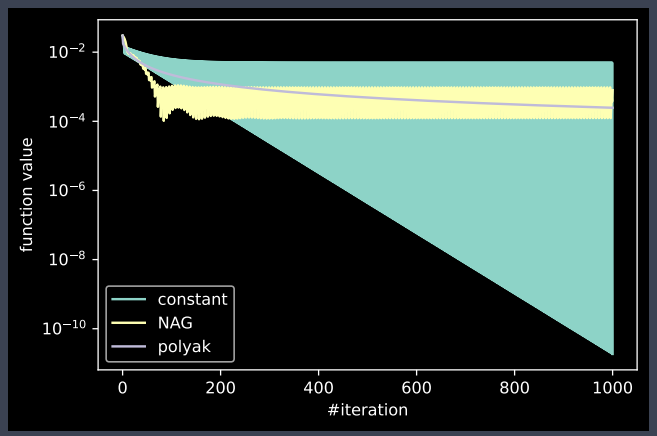
» Examples

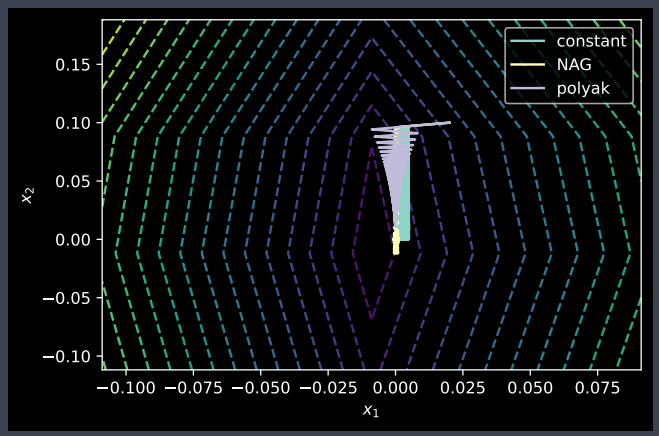
* Non-smooth function $f(x) = |x_1| + x_2^2$
* Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.005$, constant $\alpha = 0.005$





* Nesterov $\beta = (t-1)/(t+2)$, $\alpha = 0.001$, constant $\alpha = 0.005$





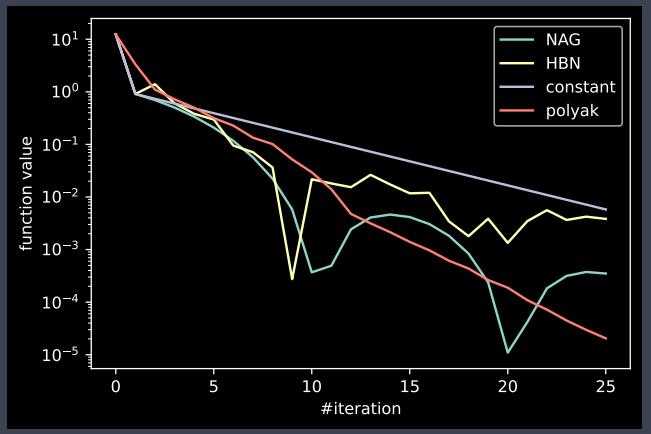
* Can tweak α to improve Nesterov, but shows oscillations/chattering after initial fast convergence

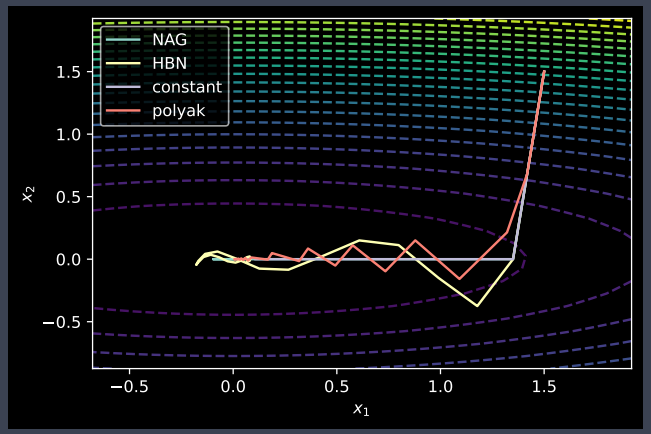
» Nesterov vs Heavy Ball Momentum?

* Two differences:
  * $\nabla f(x_t)$ vs $\nabla f(x_t + \beta z_t)$
  * $\beta = 0.9$ vs $\beta = (t-1)/(t+2)$
* Which matters (or do both matter)?
* Using $\beta = (t-1)/(t+2)$ in both Heavy Ball and Nesterov approaches it turns out that the performance of the two approaches is v similar, let's look at some examples ...

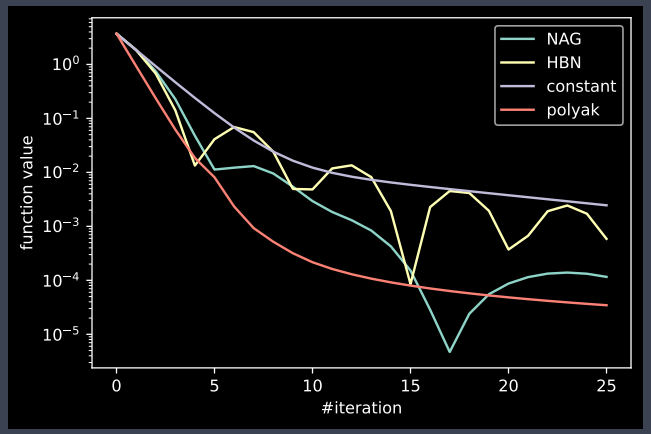
» Nesterov vs Heavy Ball Momentum?

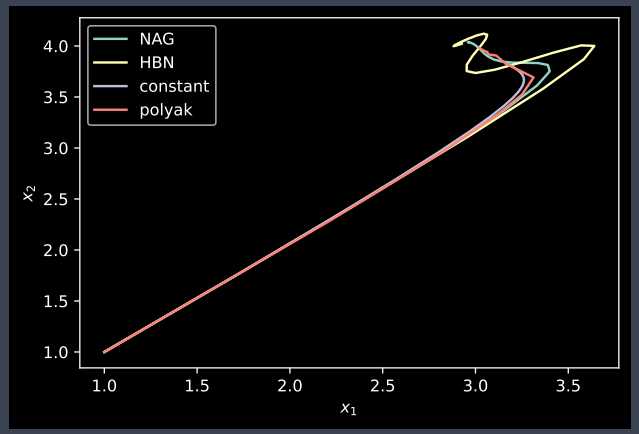
* Quadratic: Nesterov/HB $\beta = (t-1)/(t+2)$, $\alpha = 0.1$





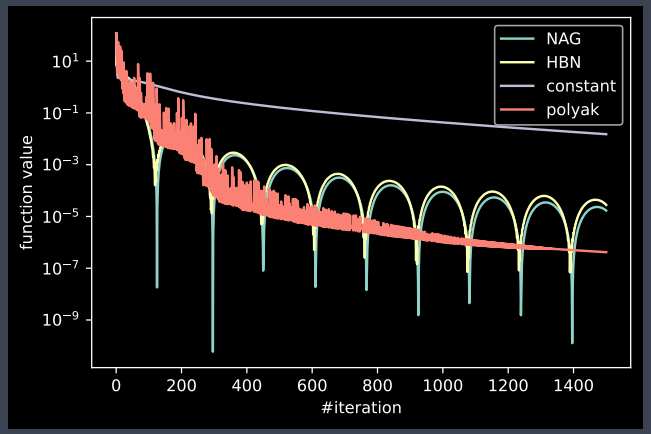
* Quadratic Loss: Nesterov/HB $\beta = (t-1)/(t+2)$, $\alpha = 0.5$

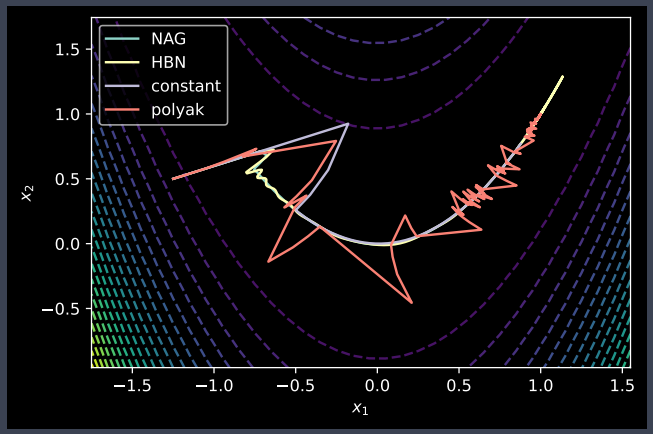




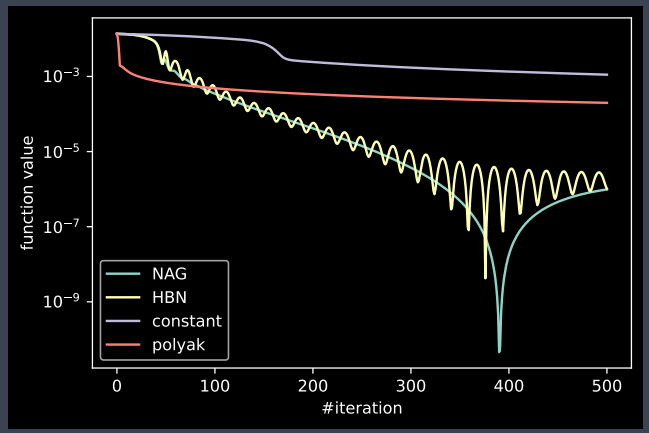
» Nesterov vs Heavy Ball Momentum?

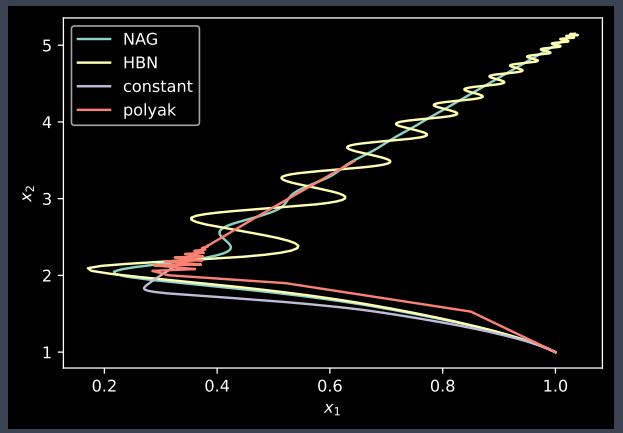
* Rosenbrock function: Nesterov/HB $\beta = (t-1)/(t+2)$, $\alpha = 0.001$





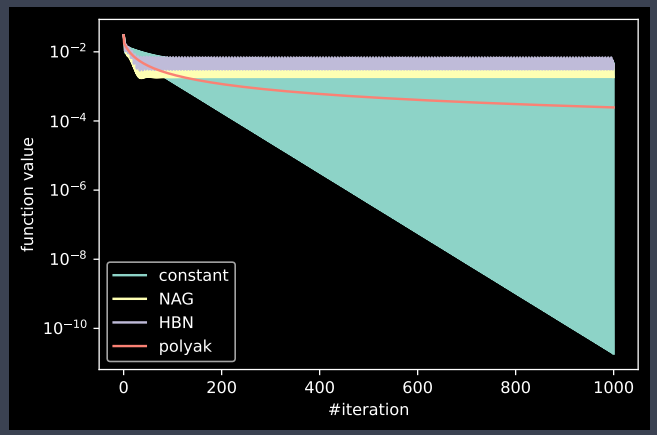
* Toy neural net loss: Nesterov/HB $\beta = (t-1)/(t+2)$, $\alpha = 0.75$

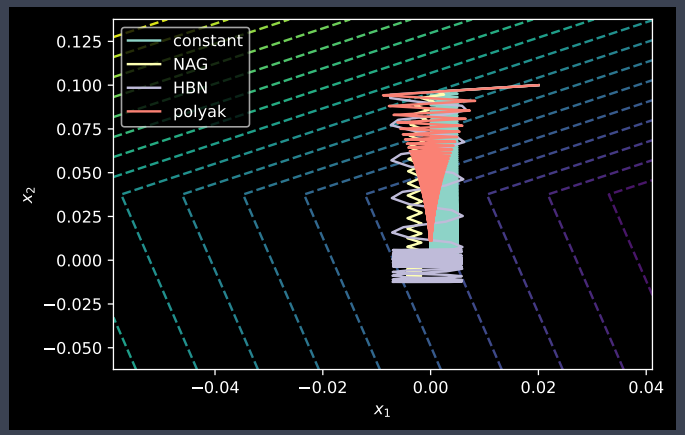




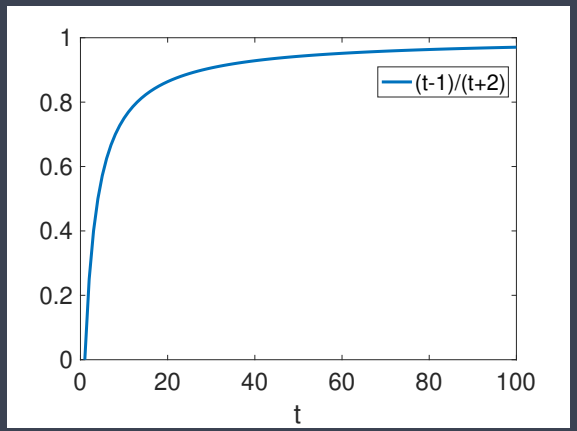
» Nesterov vs Heavy Ball Momentum?

* Non-smooth function $f(x) = |x_1| + x_2^2$
* Nesterov/HB $\beta = (t-1)/(t+2)$, $\alpha = 0.005$





» What is $\beta = (t-1)/(t+2)$ schedule doing?



- Initially β is 0, then increases over time towards 1.
- When $\beta = 0$ the update reverts to constant step size strategy, as β increases we start to use more "momentum" when choosing the change to x.
- Can choose to cap maximum value of β e.g. $\beta = min((t-1)/(t+2), 0.95)$

» Adam[1]

* Adam $\approx$ RMSprop + heavy ball

$$m_{t+1} = \beta_1 m_t + (1-\beta_1)\nabla f(x_t)$$

$$v_{t+1} = \beta_2 v_t + (1-\beta_2)[\frac{\partial f}{\partial x_1}(x_t)^2, \frac{\partial f}{\partial x_2}(x_t)^2, \ldots, \frac{\partial f}{\partial x_n}(x_t)^2]$$

$$\hat{m} = \frac{m_{t+1}}{(1-\beta_1^t)},\ \hat{v} = \frac{v_{t+1}}{(1-\beta_2^t)}$$

$$x_{t+1} = x_t - \alpha[\frac{\hat{m}}{\sqrt{\hat{v}_1}+\epsilon}, \frac{\hat{m}}{\sqrt{\hat{v}_2}+\epsilon}, \ldots, \frac{\hat{m}}{\sqrt{\hat{v}_n}+\epsilon}]$$

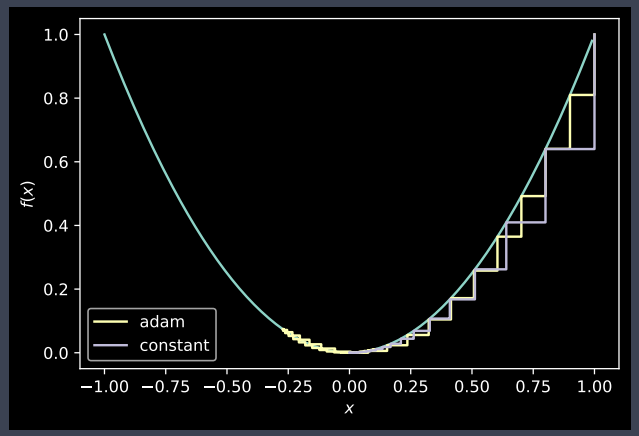
with $\nabla f(x_t) = [\frac{\partial f}{\partial x_1}(x_t), \frac{\partial f}{\partial x_2}(x_t), \ldots, \frac{\partial f}{\partial x_n}(x_t)]$

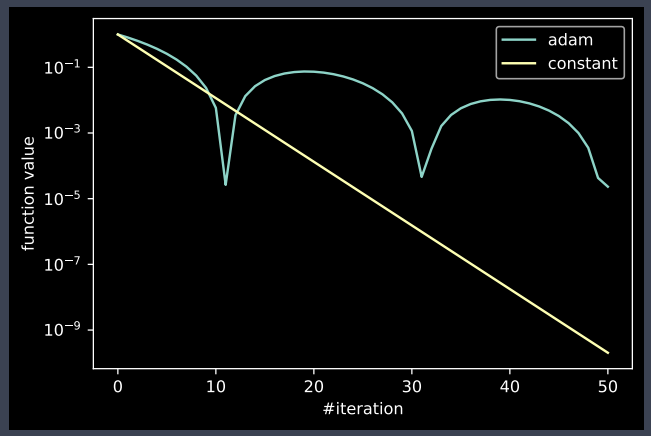
* m is running average of gradient $\nabla f(x_t)$, v is running average of square gradients
* Use different step size for each element of vector x (similarly to Adagrad and RMSprop)
* Step for i'th element of x is $\frac{\hat{m}}{\sqrt{\hat{v}_i}+\epsilon} \rightarrow \hat{m}$ is as in heavy-ball approach, $\sqrt{\hat{v}_i}$ as in RMSprop, ϵ is to avoid division by zero.
* Need to manually select β_1, β_2, α. Common choices $\beta_1 = 0.9, \beta_2 = 0.999$ (suggested by https://arxiv.org/abs/1412.6980).

[1] https://arxiv.org/abs/1412.6980

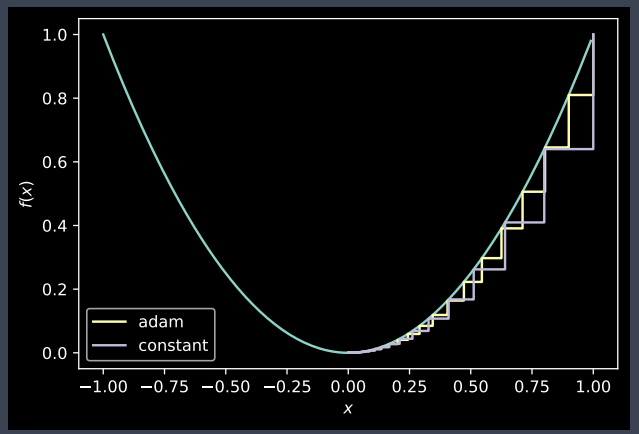
» Examples

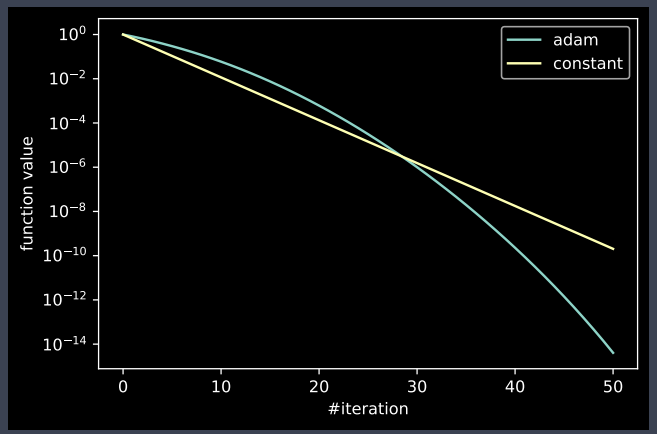
* Example: $f = x^2$, starting point $x = 1$
* Adam $\beta_1 = 0.9, \beta_2 = 0.999$, $\alpha = 0.1$, constant step $\alpha = 0.1$





* Adam $\beta_1 = 0.25, \beta_2 = 0.999$, $\alpha = 0.1$, constant step $\alpha = 0.1$

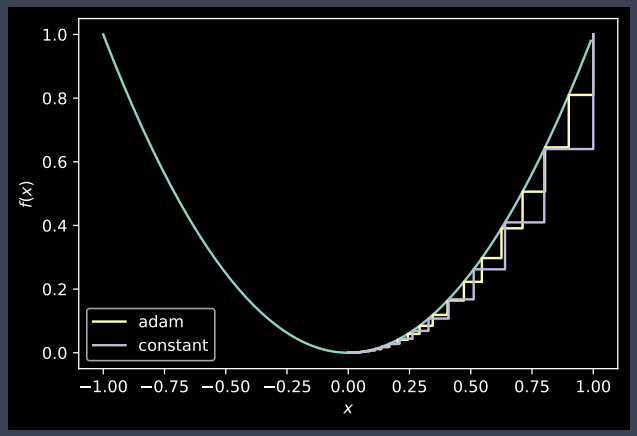


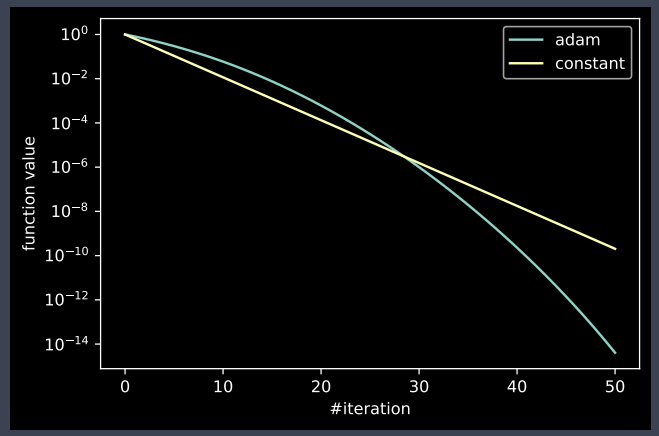


* Observe the “bumpy” behaviour of function value vs time when using momentum.
* Choice of β_1 matters, default choice $\beta_1 = 0.9$ is poor in this example

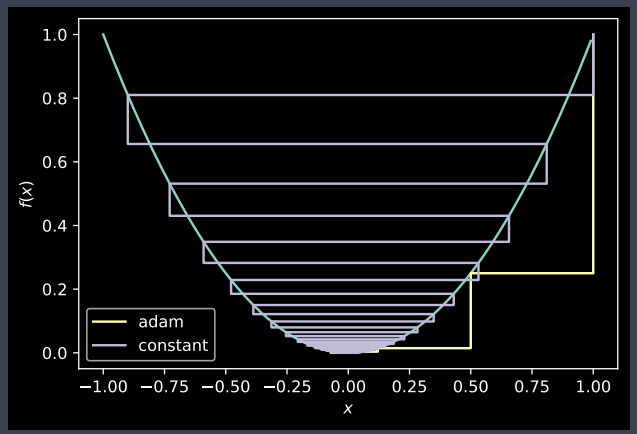
» Examples

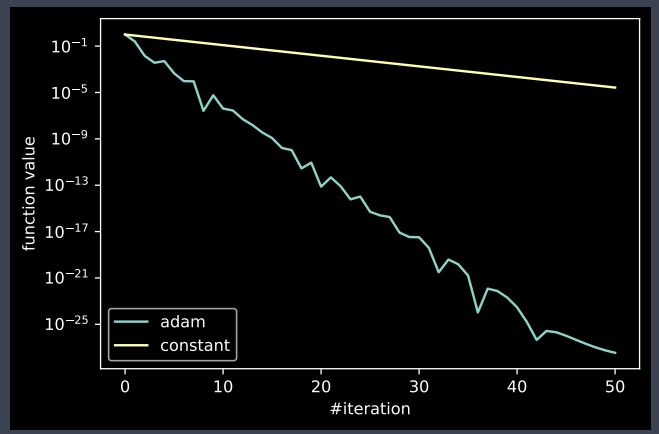
* Example: $f = x^2$, starting point $x = 1$
* Adam $\beta_1 = 0.25, \beta_2 = 0.999$, $\alpha = 0.1$, constant step $\alpha = 0.1$





* Can increase α: Adam $\alpha = 0.5$, constant step $\alpha = 0.95$

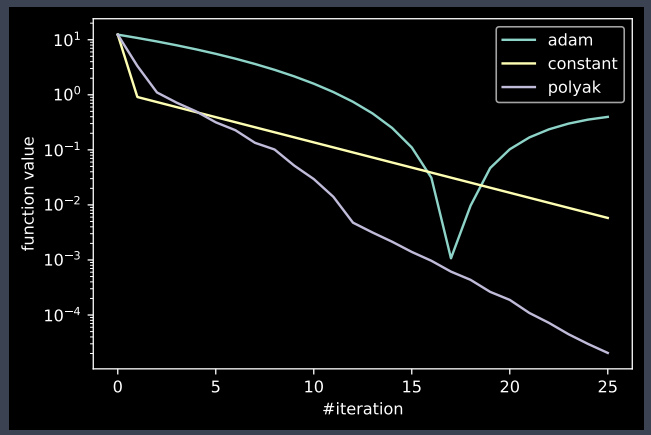


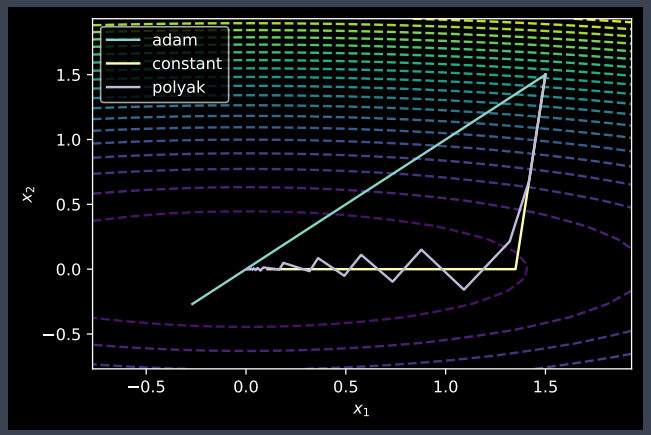


* For constant step size, $\alpha_0 > 1$ gives divergent solution, and very oscillatory solutions for $\alpha_0 > 0.9$.
* With Adam can use larger step size without destabilising solution

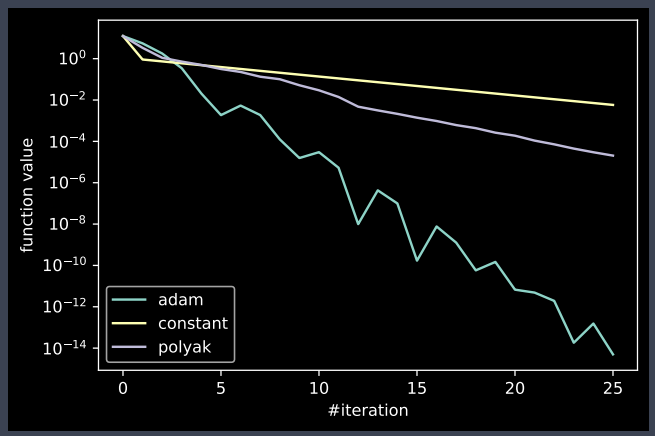
» Examples

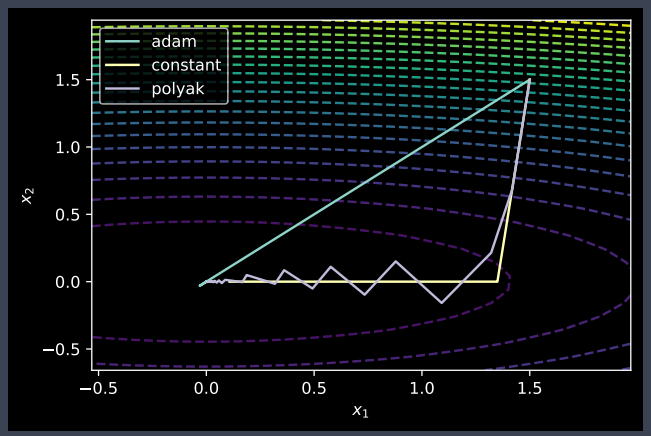
* Quadratic: Adam $\beta_1 = 0.9, \beta_2 = 0.999$, $\alpha = 0.1$, constant step $\alpha = 0.1$





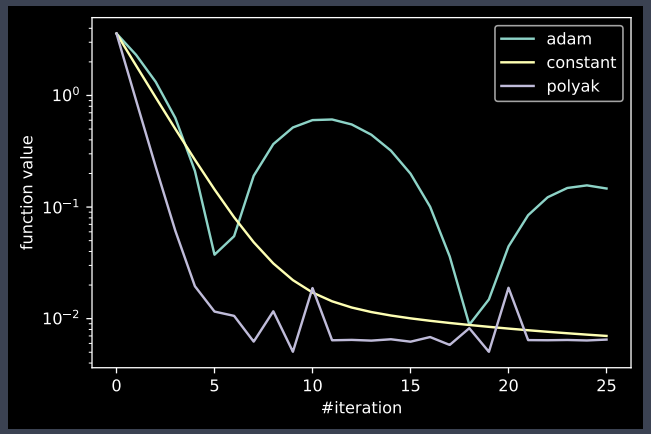
* Decrease β_1, increase α: Adam $\beta_1 = 0.25, \beta_2 = 0.999$, $\alpha = 0.5$ (increasing α causes solution to diverge)„ constant step $\alpha = 0.1$

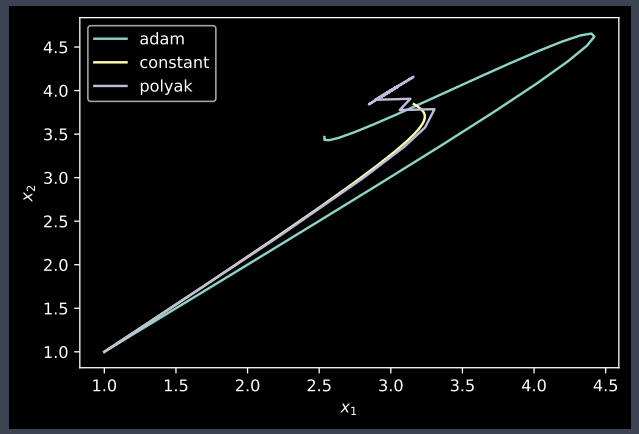




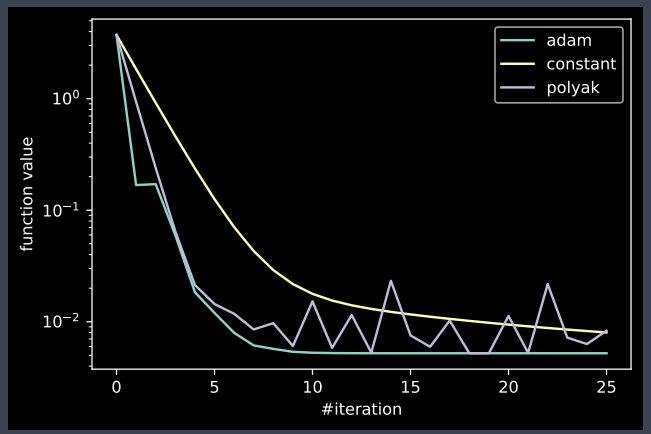
» Examples

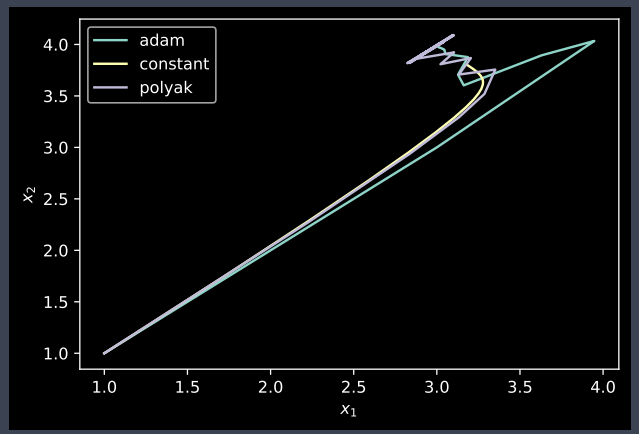
* Quadratic Loss: Adam $\beta_1 = 0.9, \beta_2 = 0.999$, $\alpha = 0.5$, constant step $\alpha = 0.5$





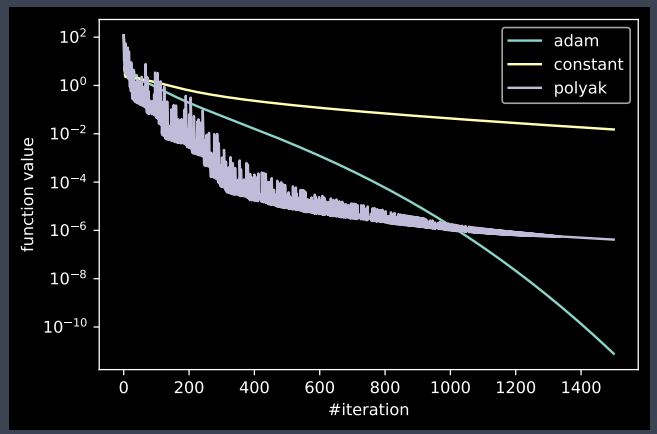
* Decrease β_1, increase α: Adam $\beta_1 = 0.25, \beta_2 = 0.999$, $\alpha = 2$ (increasing α causes solution to diverge)„ constant step $\alpha = 0.5$

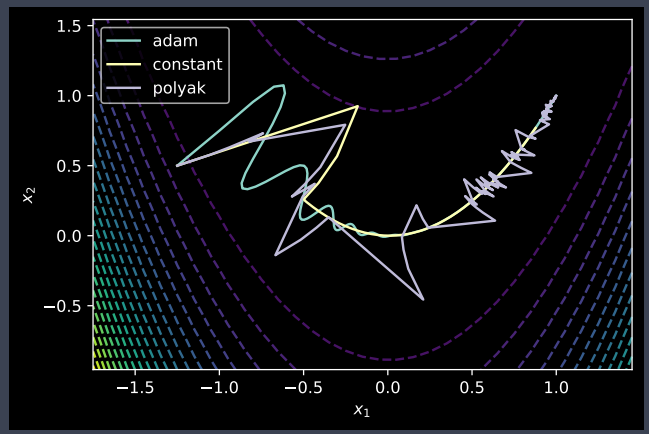




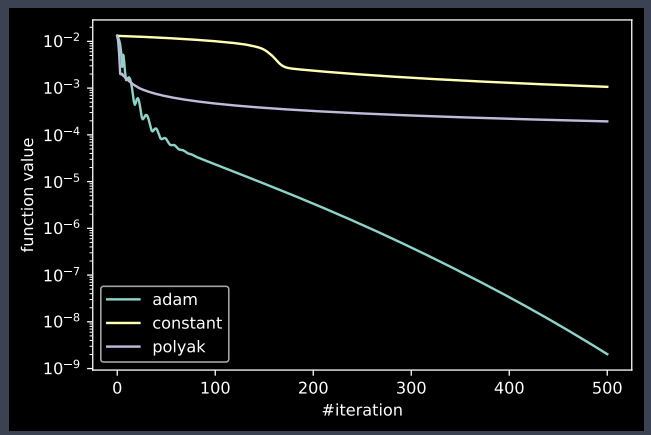
» Examples

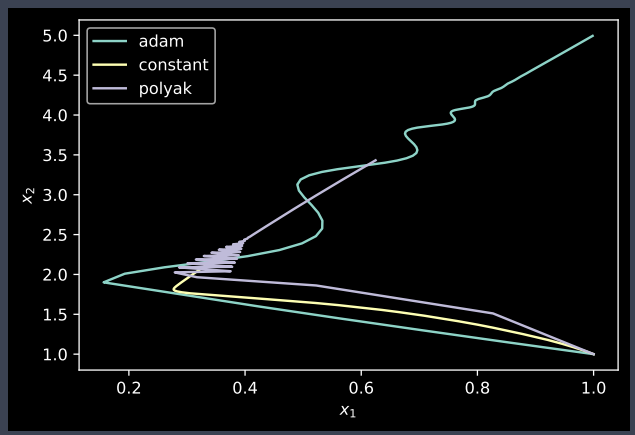
* Rosenbrock function: Adam $\beta_1 = 0.9, \beta_2 = 0.999, \alpha = 0.25$ (increasing α causes solution to diverge), constant step $\alpha = 0.002$





* Toy neural net loss: Adam $\beta_1 = 0.9, \beta_2 = 0.999, \alpha = 0.5$, constant step $\alpha = 0.5$

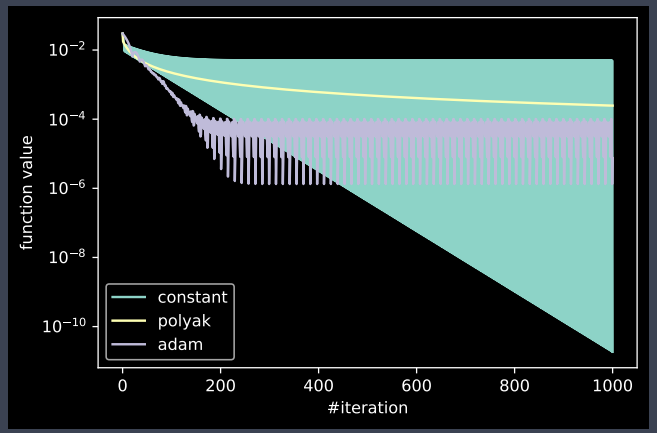


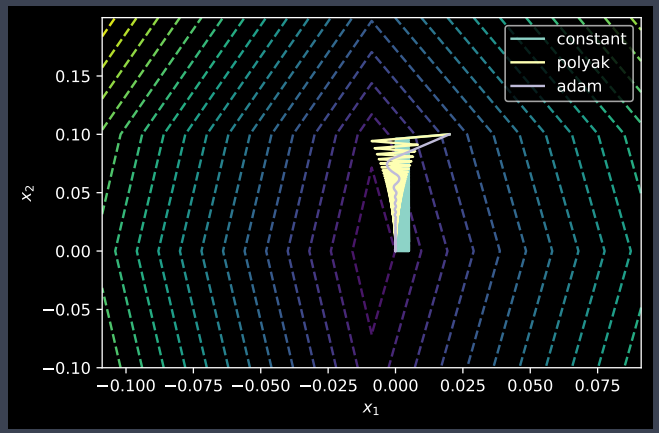


* Adam convergence rate similar to Polyak for Rosenbrock function, faster for toy neural net.

» Examples

* Non-smooth function $f(x) = |x_1| + x_2^2$
* Adam $\beta_1 = 0.9, \beta_2 = 0.999$, $\alpha = 0.001$, constant step $\alpha = 0.005$





* Adam oscillations/chattering after initial fast convergence
* Note: Adam α here is smaller than for constant step size (increasing Adam α increaes oscillations and slows convergence)

» Summary

- When use $\beta = (t-1)/(t+2)$ schedule, Heavy Ball and Nesterov momentum approaches are pretty similar in our examples
- Nesterov/Heavy Ball can accelerate convergence, but still need to tune α.
- Adam $\approx$ RMSpop + Heavy Ball. Can accelerate convergence but need to tune β_1 and α (suggest default values are *not* universal).