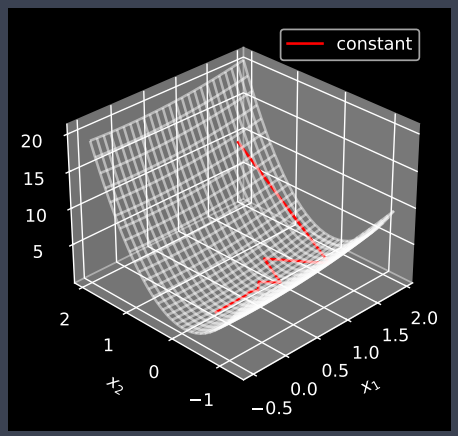
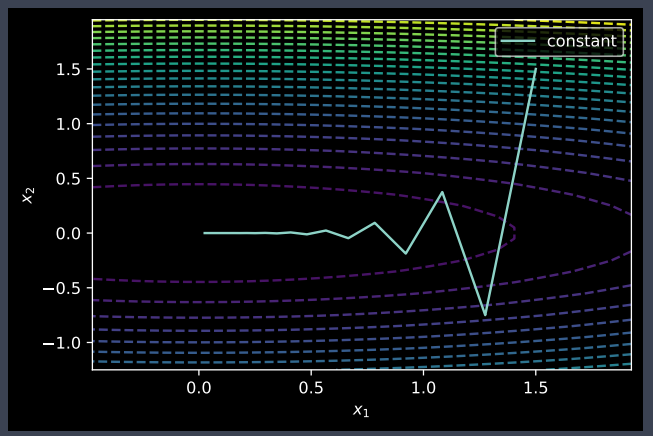
» Momentum Methods: Motivation

* In steep, narrow valleys moving in direction of steepest descent might not be best. For example:
* $f(x) = x_1^2 + 10x_2^2$, starting value $x = [1.5, 1.5]$, constant step size $alpha = 0.15$

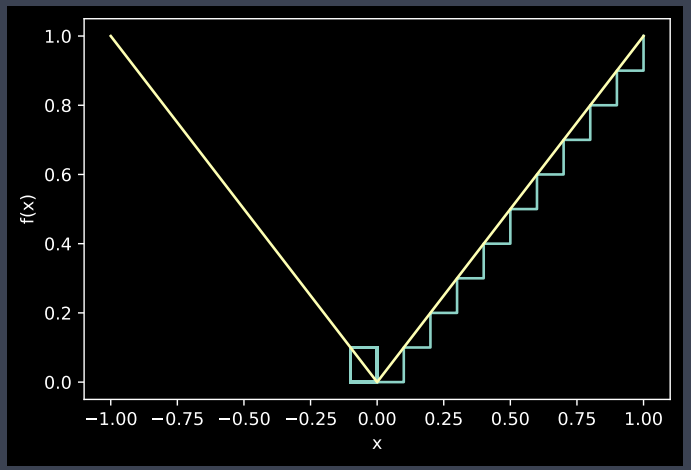


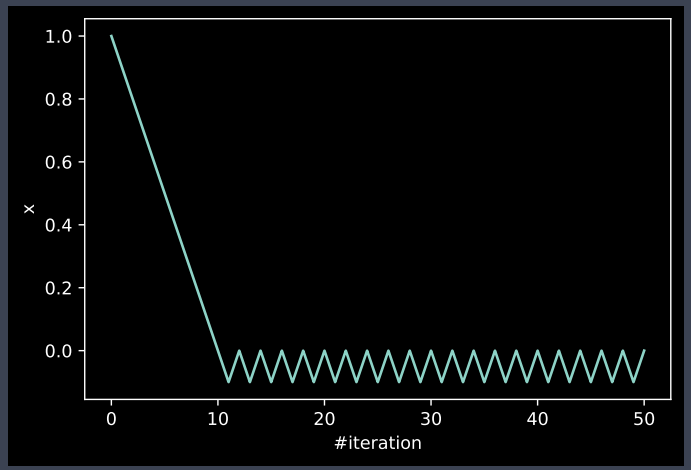


* Wouldn't it be faster to take a less zig-zag path towards minimum?
* What if we took the average of the zig-zag steps?

» Momentum Methods: Motivation

* Oscillations or “chattering” are common when function is non-smooth. For example:
* $f(x) = |x|$, starting value $x0 = 1$, constant step size $alpha = 0.1$





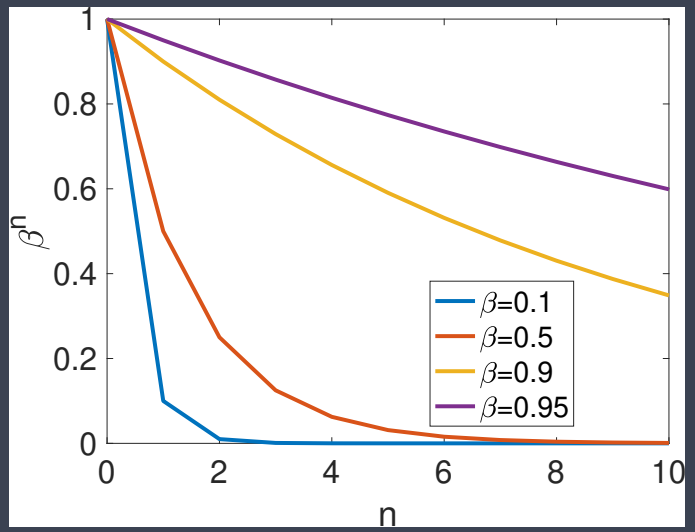
* Function decreases until it reaches vicinity of minimum, then oscillates forever - never converges to minimum.
* What if we averaged the oscillations and used that value for x?

» Weighted Sum/Running Average

* $z_{t+1} = \beta z_t + u_t$ with $z_0 = 0$, $0 < \beta < 1$
* Expanding this out:

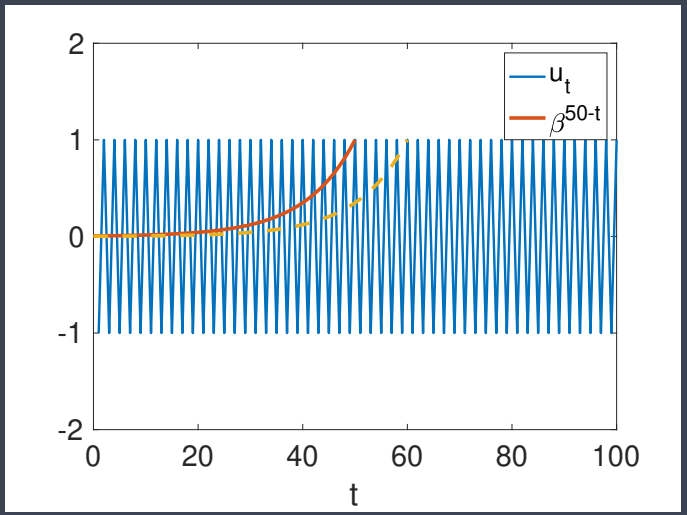
$$z_1 = u_0$$
$$z_2 = \beta z_1 + u_1 = \beta u_0 + u_1$$
$$z_3 = \beta z_2 + u_2 = \beta^2 u_0 + \beta u_1 + u_2$$

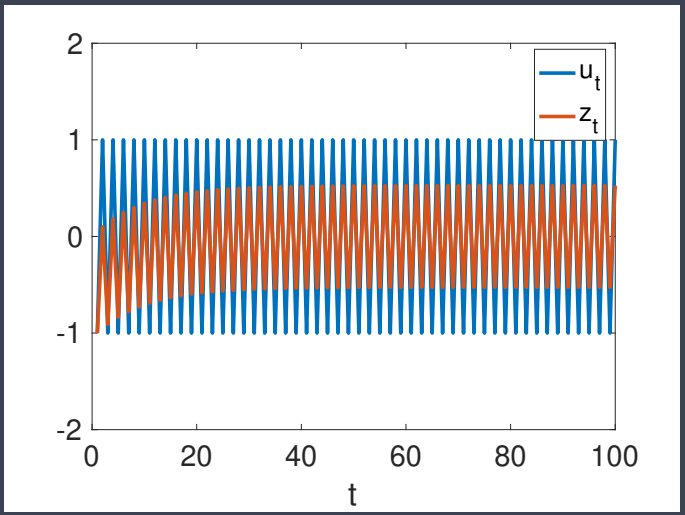
so z_t is weighted sum of $u_0, u_1, \ldots, u_{t-1}$. E.g. with $\beta = 0.9$ then $\beta = 0.9$, $\beta^2 = 0.81$, $\beta^3 = 0.729$ etc



» Weighted Sum/Running Average

* Example: $\beta = 0.9$, $u_t = (-1)^t$ i.e. $u_0 = 0, u_1 = 1, u_2 = +1, u_3 = -1, \ldots$
* $z_{t+1} = \beta z_t + u_t$, $z_0 = 0$:

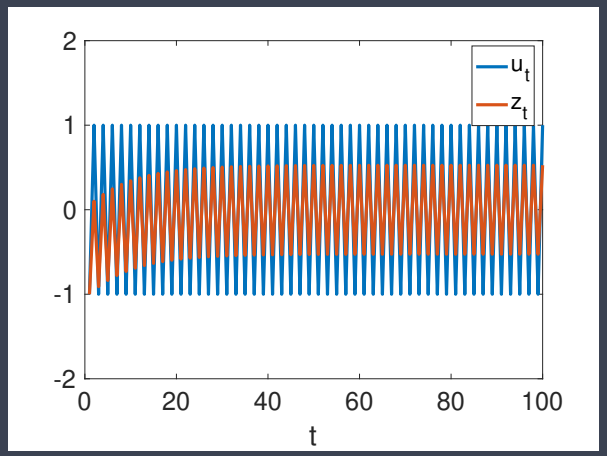




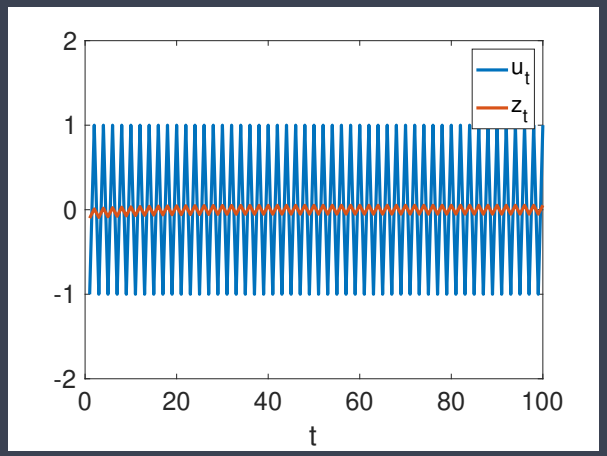
* z_t tends to smooth out (or average) the zig-zags

» Weighted Sum/Running Average

* Example: $\beta = 0.9$, $u_t = (-1)^t$ i.e. $u_0 = 0, u_1 = 1, u_2 = +1, u_3 = -1, \ldots$



$z_{t+1} = \beta z_t + u_t$



$z_{t+1} = \beta z_t + (1 - \beta) u_t$

* $1 + \beta + \beta^2 + \cdots = 1/(1 - \beta)$, so often scale z_t by $1 - \beta$ i.e. use $z_{t+1} = \beta z_t + (1 - \beta) u_t$
* Another way to think about this: suppose $u_t =$ *constant* u and $z_t =$ *constant* z .
  * $z = \beta z + (1 - \beta) u$ rearranges to $(1 - \beta) z = (1 - \beta) u$ i.e. $z = u$.
  * $z = \beta z + u$ rearranges to $(1 - \beta) z = u$ i.e. $z = u/(1 - \beta)$.

» Polyak Momentum/Heavy Ball

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

for k in range(num_iters):

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

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

$t = t + 1$

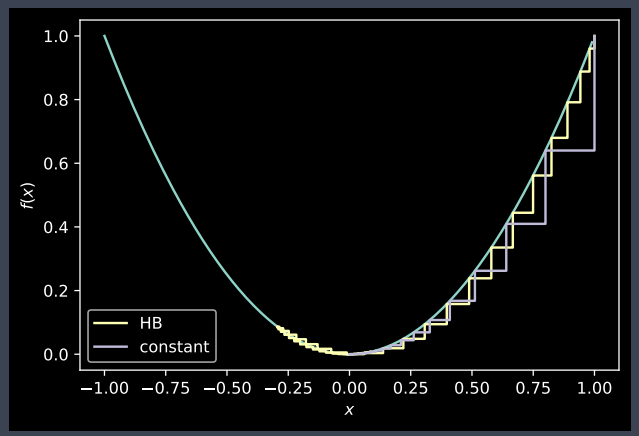
* Step z_t at iteration t is weighted sum of past gradients $\alpha \nabla f$:

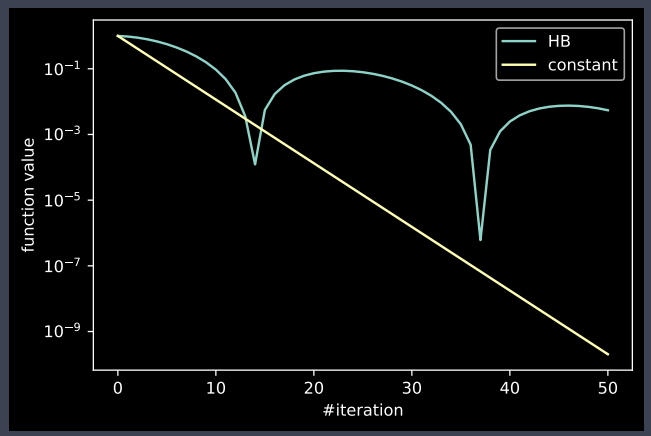
$$z_{t+1} = \beta^{t-1} \alpha \nabla f(x_0) + \beta^{t-2} \alpha \nabla f(x_1) + \cdots + \alpha \nabla f(x_t)$$

* 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

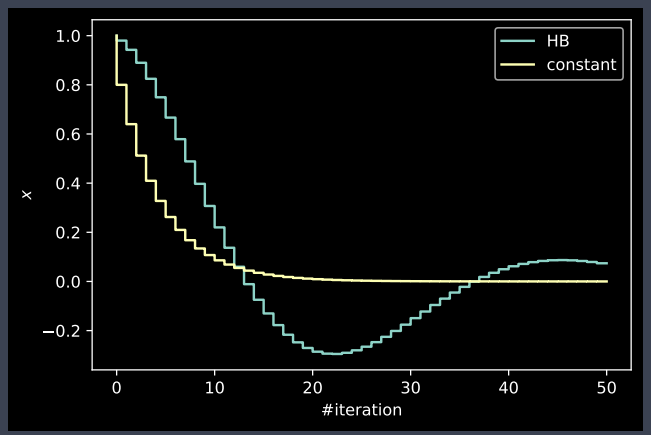
» Too Much Momentum

* Example: $f = x^2$, starting point $x = 1$
* HB $\beta = 0.9$, $\alpha_0 = 0.01$, constant $\alpha_0 = 0.1$ (so HB α_0 is $1 - \beta$ times constant α_0)



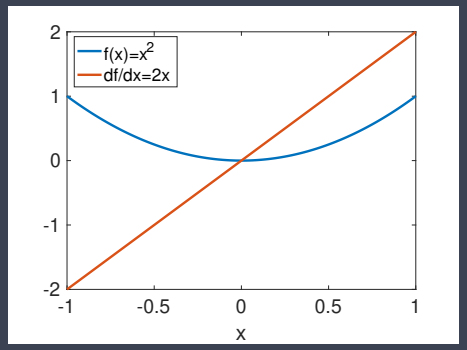


* Adding momentum seems ot have introduced new oscillations! What's going on here ...
* Hint: see that x overshoots (goes past) the minimum at $x = 0$

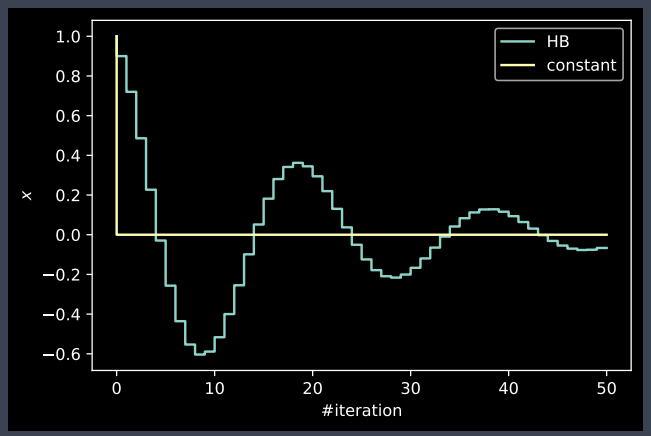


» Too Much Momentum

* Example: $f(x) = x^2$, $\frac{df}{dx}(x) = 2x$



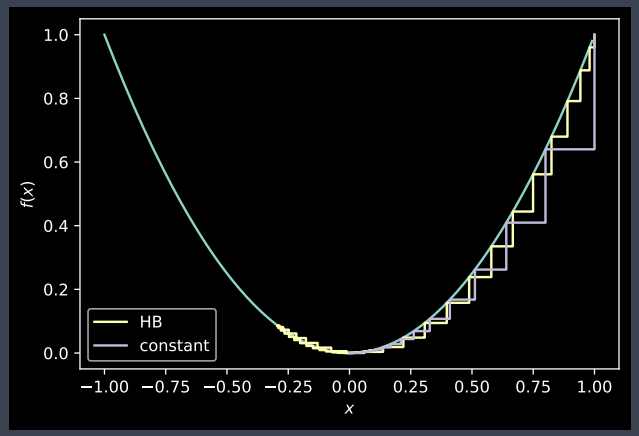
* Step at iteration t is $z_{t+1} = \beta^{t-1}\alpha\frac{df}{dx}(x_0) + \beta^{t-2}\alpha\frac{df}{dx}(x_1) + \cdots + \alpha\frac{df}{dx}(x_t)$

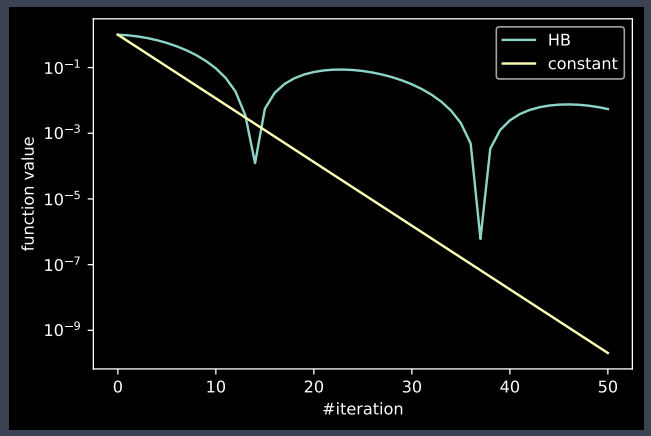


* $x \approx 0$ around #iteration 12, but HB step size is high, so shoot right past the minimum.
* x then becomes negative, and so does df/dx. HB step size starts to decrease $\rightarrow$ *but takes a while for step to become negative due to momentum/averaging, so x overshoots*
* $x \approx 0$ around #iteration 38, HB step size again big, so overshoot

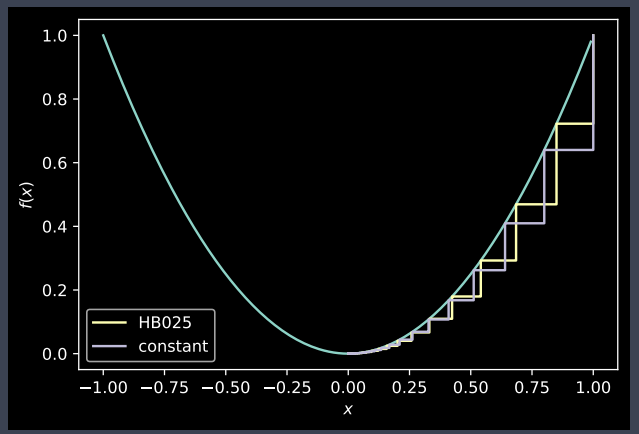
» Examples

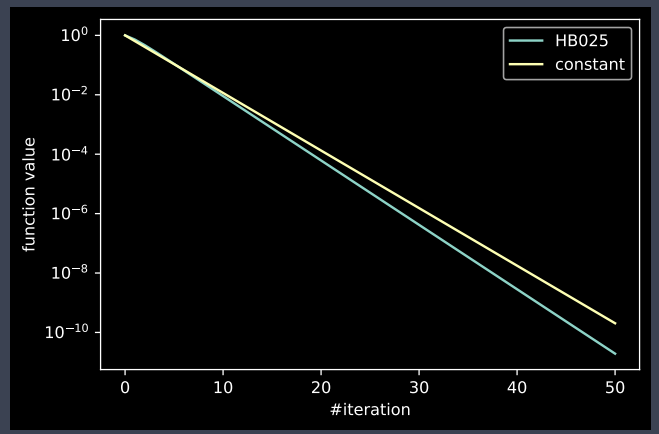
* Example: $f = x^2$, starting point $x = 1$
* HB $\beta = 0.9$, $\alpha = 0.01$, constant $\alpha_0 = 0.1$





* HB $\beta = 0.25$, $\alpha = 0.075$, constant $\alpha_0 = 0.1$

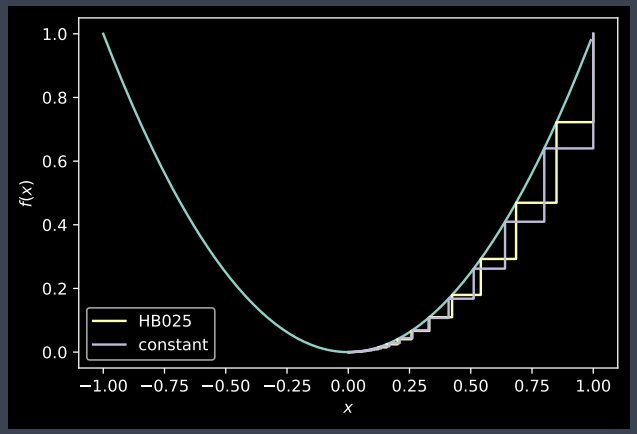


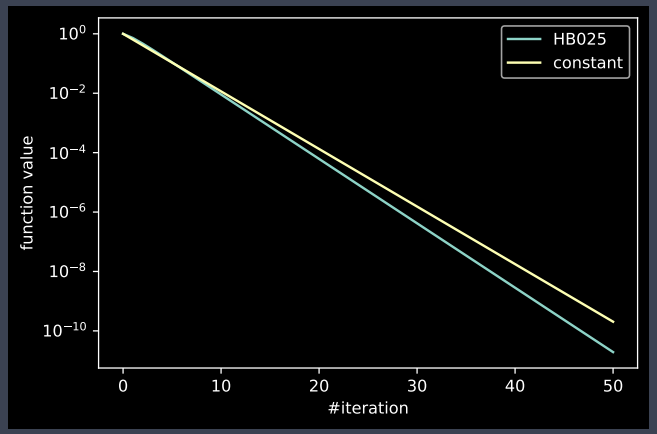


* Decreasing β (reducing momentum i.e. weight given to past gradients) removes the oscillations.

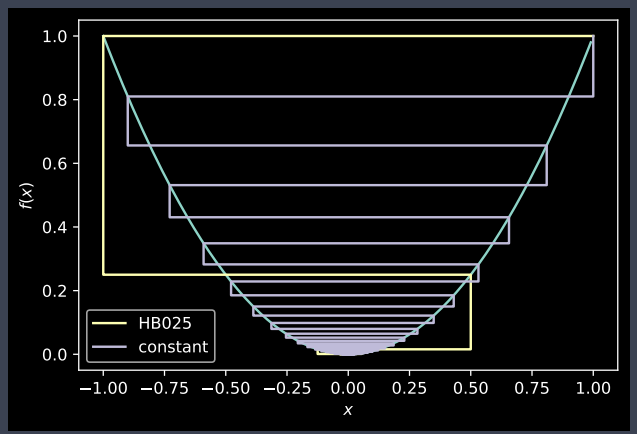
» Examples

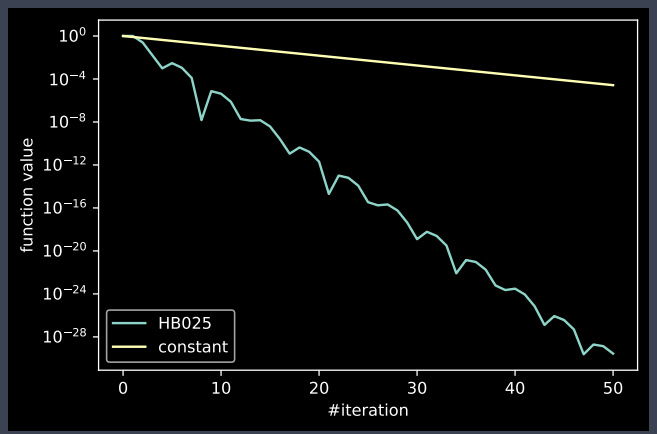
* Example: $f = x^2$, starting point $x = 1$
* HB $\beta = 0.25$, $\alpha = 0.075$, constant $\alpha_0 = 0.1$





* HB $\beta = 0.25$, $\alpha = 1$, constant $\alpha_0 = 0.95$

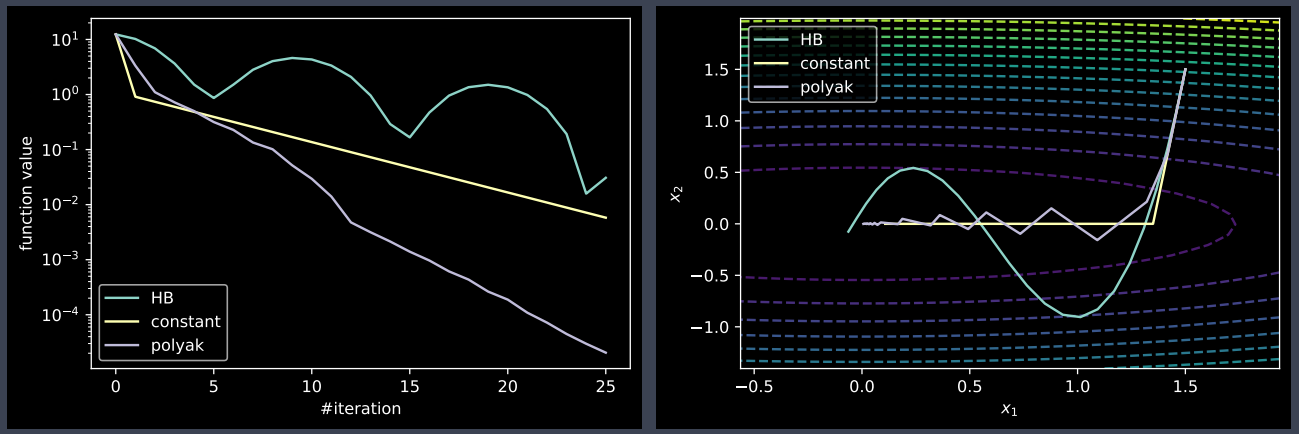




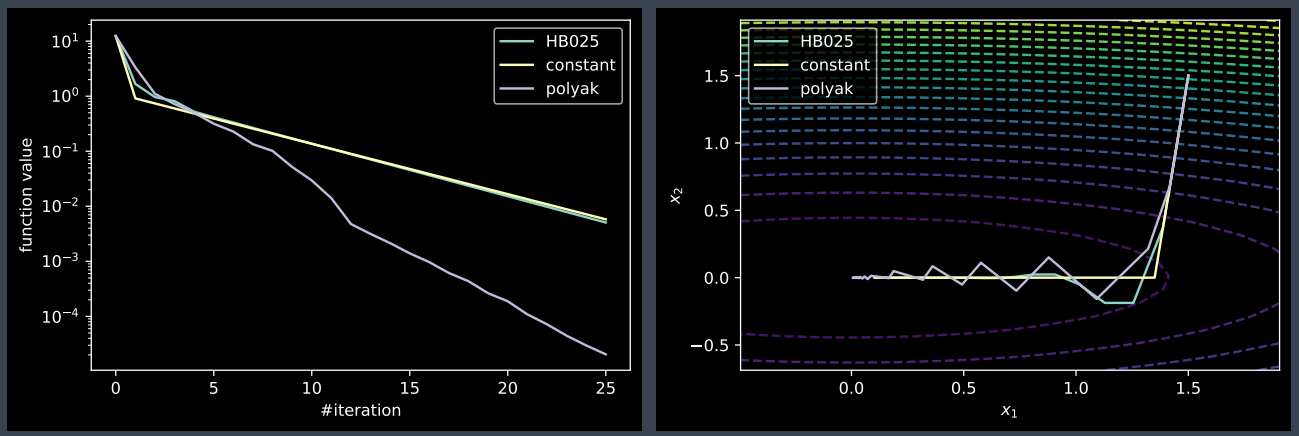
* For constant step size, $\alpha_0 > 1$ gives divergent solution, and very oscillatory solutions for $\alpha_0 > 0.9$.
* With HB can use larger step size without destabilising solution $\rightarrow$ faster convergence

» Examples

* Quadratic:
* HB $\beta = 0.9$, $\alpha = 0.01$, constant $\alpha_0 = 0.1$



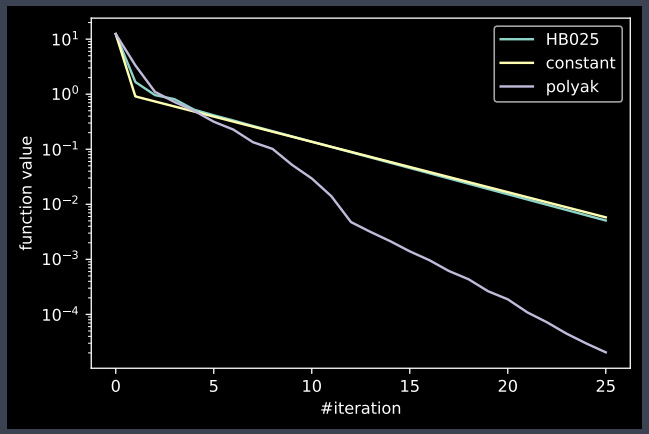
* HB $\beta = 0.25$, $\alpha = 0.075$, constant $\alpha_0 = 0.1$

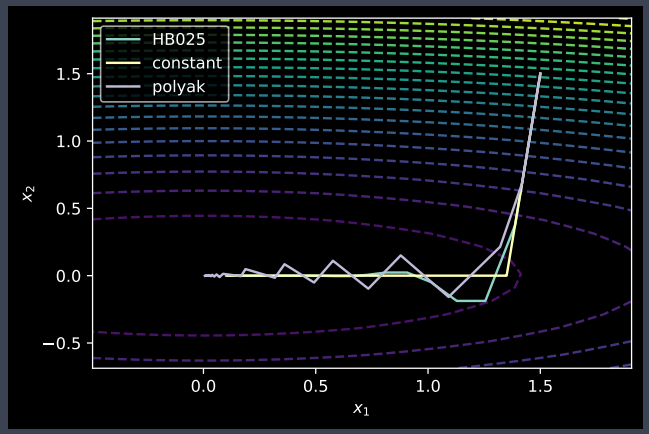


* Again, HB overshoots minimum when $\beta = 0.9$ and *increases* oscillations. Using less momentum $\beta = 0.25$ removes oscillations $\rightarrow$ performance no better than constant step size strategy but with HB can increase step size

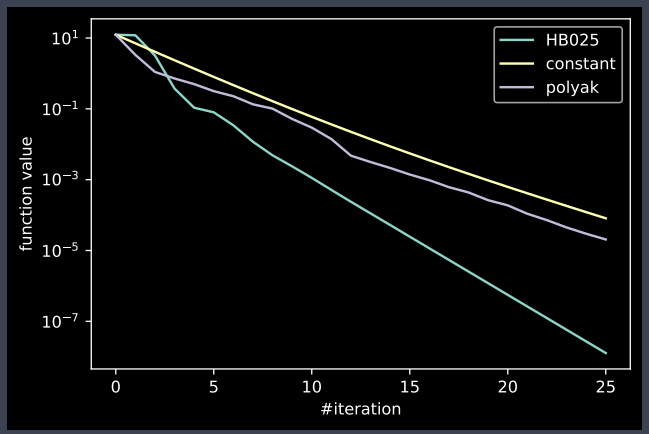
» Examples

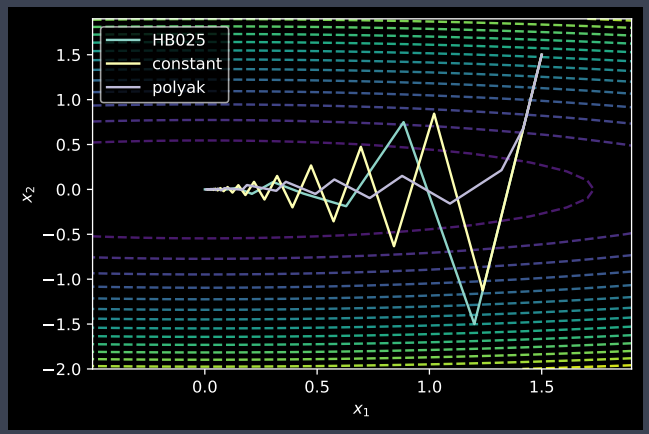
* Quadratic:
* HB $\beta = 0.25$, $\alpha = 0.075$, constant $\alpha_0 = 0.1$





* Increase step size: HB $\beta = 0.25$, $\alpha = 0.2$, constant $\alpha_0 = 0.175$

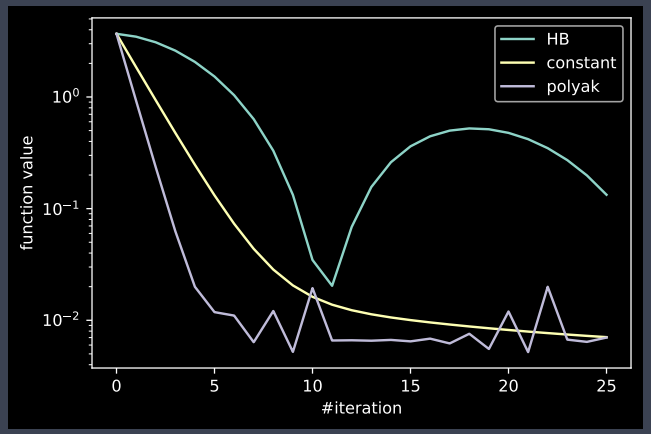


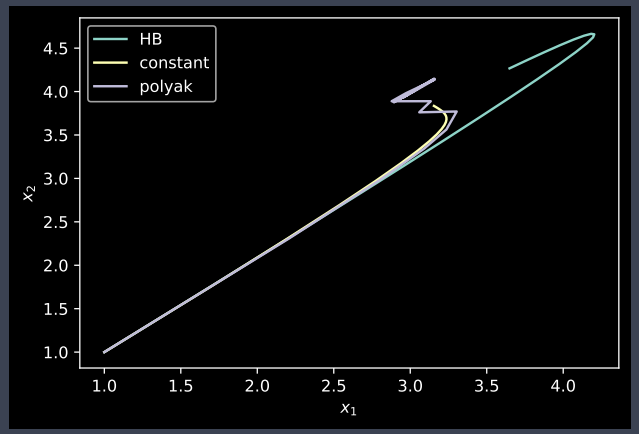


* Increasing constant step size α_0 to 0.2 causes solution to diverge

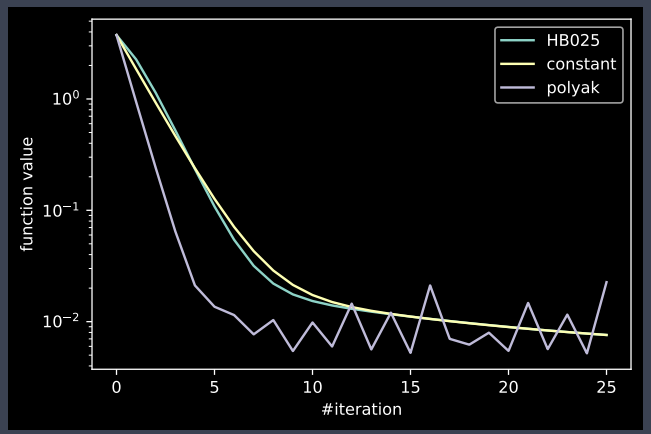
» Examples

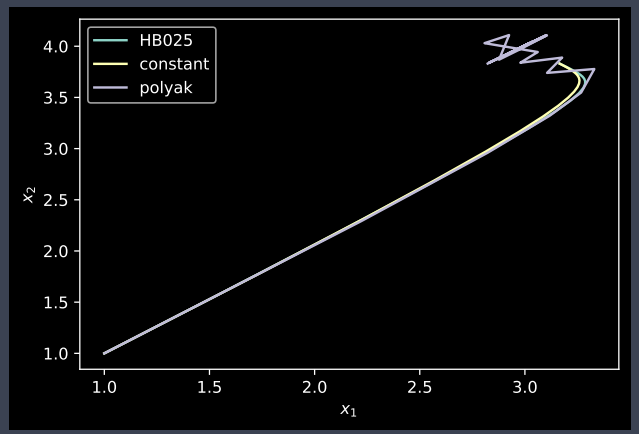
* Quadratic loss:
* HB $\beta = 0.9$, $\alpha = 0.05$, constant $\alpha_0 = 0.5$





* HB $\beta = 0.25$, $\alpha = 0.375$, constant $\alpha_0 = 0.5$

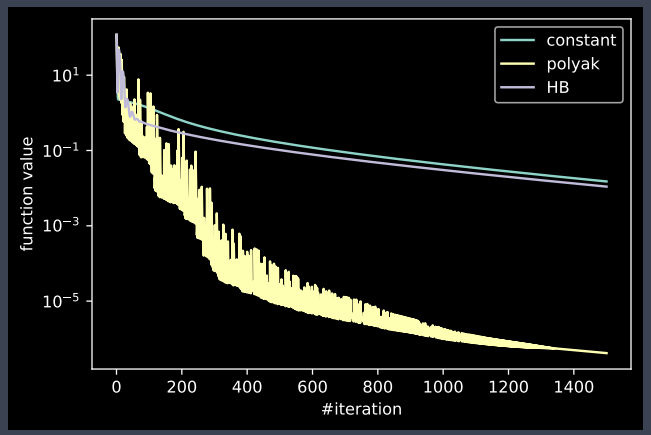


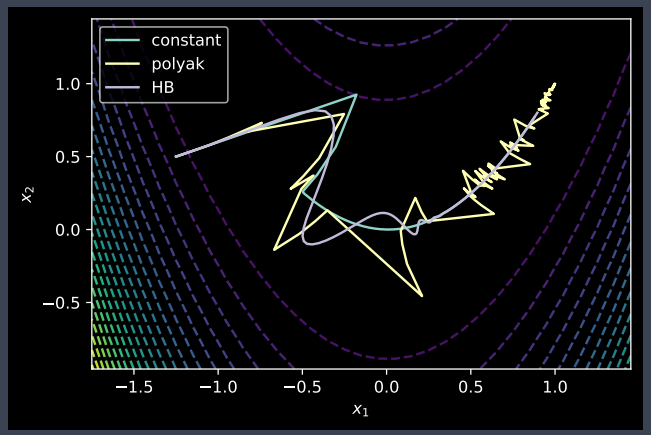


* Similarly to last example: $\beta = 0.9$ increases oscillations, $\beta = 0.25$ offers little benefit but wit HB can increase step size

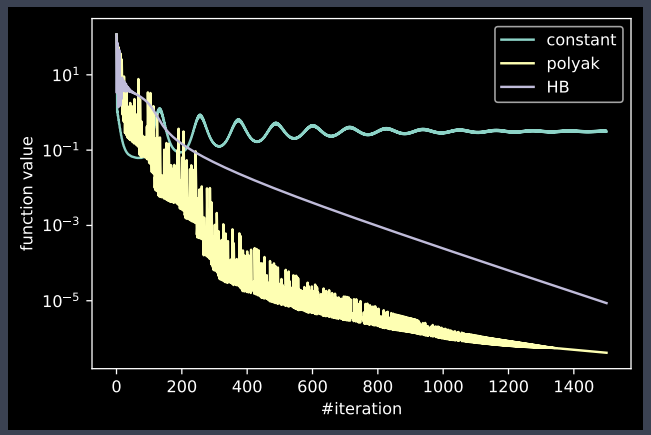
» Examples

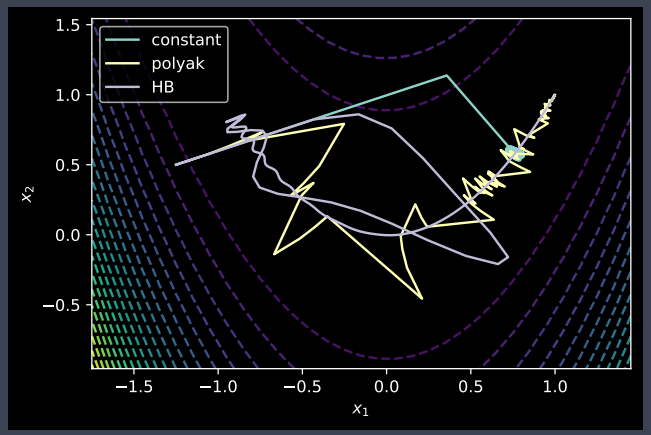
* Rosenbrock function:
* HB $\beta = 0.9$, $\alpha = 0.0002$, constant $\alpha_0 = 0.002$





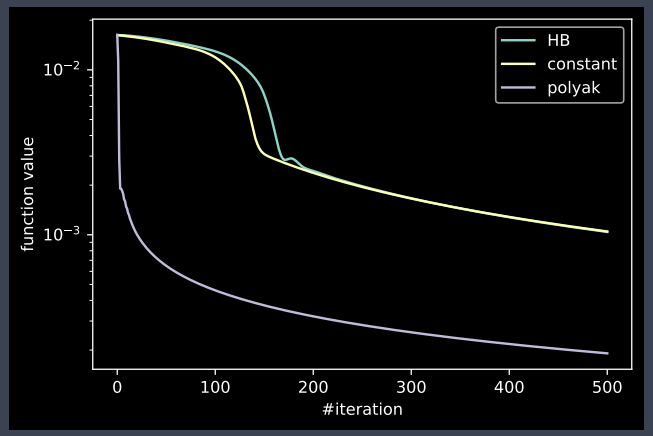
* Finally an example where $\beta = 0.9$ doesn't increase oscillations.
* Performance of HB is similar to constant step size when HB $\alpha = 0.0002$ is $(1 - \beta)$ times constant $\alpha_0 = 0.002$
* HB allows use of larger α without destabilising solution. HB $\beta = 0.9$, $\alpha = 0.0003$, constant $\alpha_0 = 0.008$

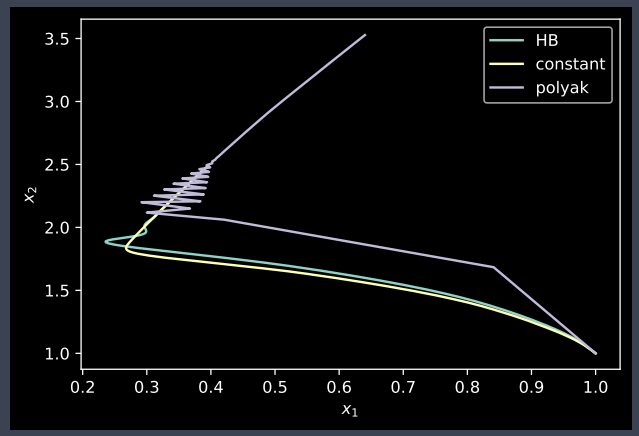




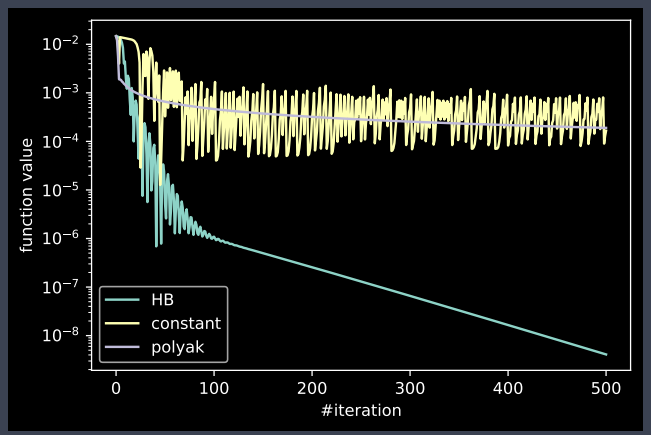
» Examples

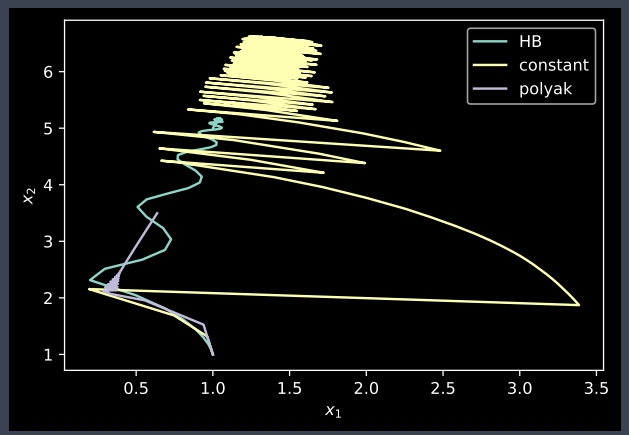
* Toy neural net loss:
* HB $\beta = 0.9$, $\alpha = 0.075$, constant $\alpha_0 = 0.75$





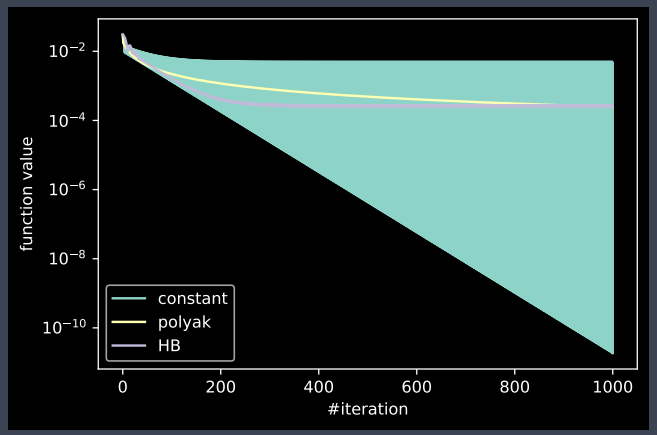
* Again HB allows larger α than constant step strategy e.g. HB $\beta = 0.9$, $\alpha = 7.5$, constant $\alpha_0 = 75$

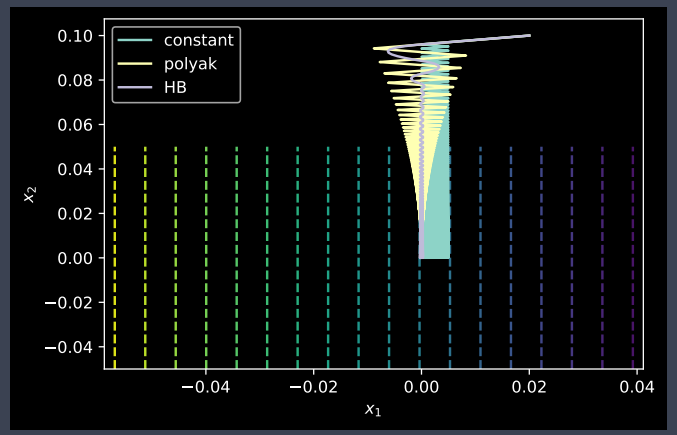




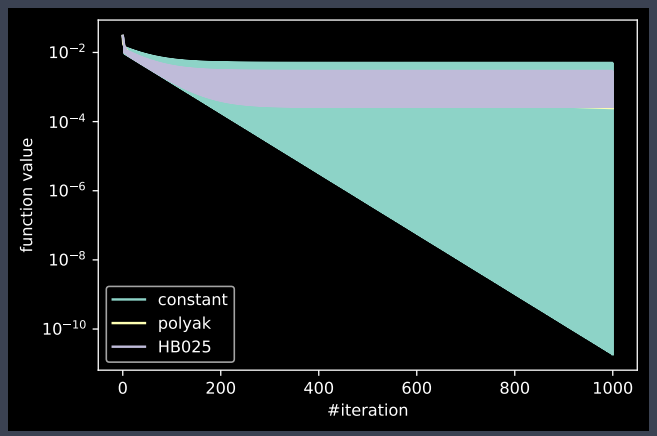
» Examples

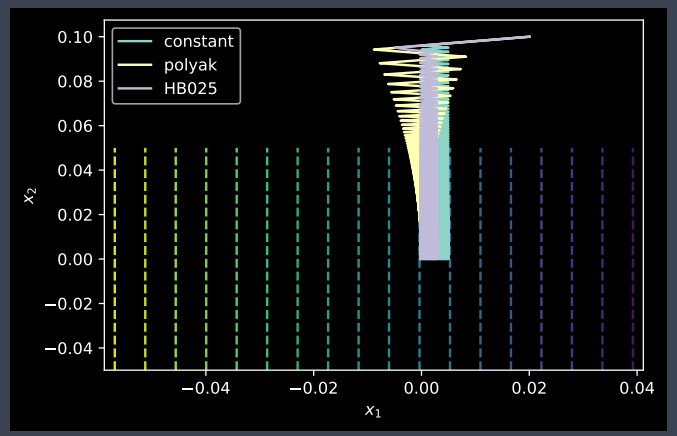
* Non-smooth function $f(x) = |x_1| + x_2^2$
* HB $\beta = 0.9$, $\alpha = 0.0005$, constant $\alpha_0 = 0.005$





* HB $\beta = 0.25$, $\alpha = 0.004$, constant $\alpha_0 = 0.005$





* Higher momentum ($\beta = 0.9$) reduces oscillations, as expected
* In this example can't increase HB α further without destabilising, but already faster than constant step size strategy

» Summary

* Use of momentum can reduce oscillations/chattering
* But need to manually tune momentum parameter β, a poor choice can *increase* oscillations/chattering
* Also need to manually tune α
* HB can allow use of a larger step size α without causing solution to diverge $\rightarrow$ faster convergence than with constant step size strategy