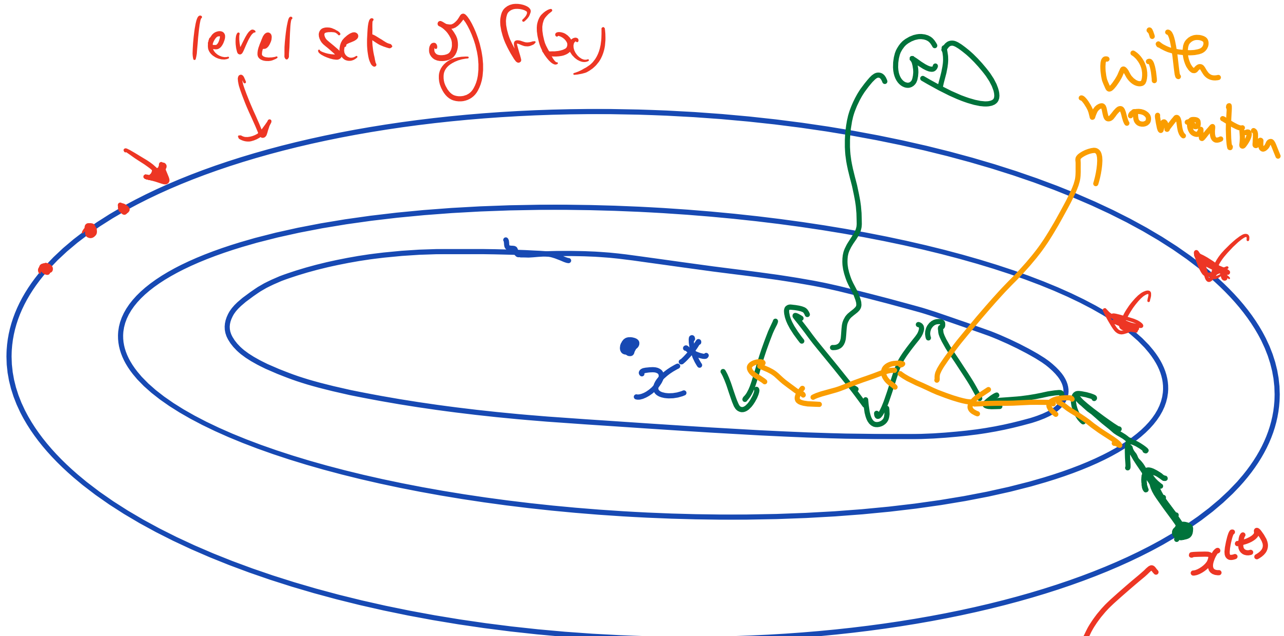
Accelerating Gradient descent:

- GD with momentum
- GD with acceleration
  Nesterov's.

Suppose f has the following level sets:



"artist's" rendition of what GD steps might look like

"Too many oscillations"

Consider the following idea:

$$\boxed{x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)}) + \underbrace{\beta(x^{(t)} - x^{(t-1)})}_{\text{"momentum term"}}}$$

"momentum term"
Change vector from the last iteration

"Intuition": If $-\nabla f(x^{(t)})$ happens to be in the same direction as $x^{(t)} - x^{(t-1)}$ (the previous step) move a little further in that direction.
Otherwise, if they are in opposite directions, move less far in those directions.

Remarks:
- This method is also known as the "heavy ball method"
- Also known as "Polyak Momentum".

We will not analyze this method in detail, but we'll do one or two illustrative examples to get a better idea of its performance:

Consider $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = \frac{\lambda}{2} x^2$$

Here momentum gives

$$x^{(t+1)} = x^{(t)} - \underbrace{\mu \lambda x^{(t)}}_{\mu \nabla f(x^{(t)})} + \beta\left(x^{(t)} - x^{(t-1)}\right)$$

$$= (1 + \beta - \lambda\mu)\, x^{(t)} - \beta x^{(t-1)}$$

$$\begin{bmatrix} x^{(t+1)} \\ x^{(t)} \end{bmatrix} = \underbrace{\begin{bmatrix} 1+\beta-\lambda\mu & -\beta \\ 1 & 0 \end{bmatrix}}_{\text{call this } M,} \begin{bmatrix} x^{(t)} \\ x^{(t-1)} \end{bmatrix}$$

M does not depend on t

For this example M is really important as it governs how the system evolves

Turns out that M has eigenvalues

$$\frac{(1+\beta-\mu\lambda) \pm \sqrt{(1+\beta-\mu\lambda)^2 - 4\beta}}{2}$$

which after some analysis (beyond the scope of this course)

implies

$$\left(x^{(t+1)}\right)^2 \leq (\text{Junk})\, \beta^t$$

- So momentum converges at a rate of β^t to the sol'n in this example.

- Same analysis extends to higher dimensional quadratics but is more complicated.

Comparing GD & momentum for general quadratics:

Let's start with GD and

$$f(x) = \frac{1}{2} x^T A x \qquad (x^* = 0)$$

where A is a symmetric PSD matrix.

Then GD would perform the iterations:

$$x^{(t+1)} = x^{(t)} - \mu A x^{(t)}$$
$$= (I - \mu A) x^{(t)} \quad \text{—— } (1)$$

How does GD converge in this case?

$$(1) \Rightarrow \quad x^{(t+1)} = (I - \mu A) x^{(t)}$$
$$= (I - \mu A)^2 x^{(t-1)}$$
$$= \cdots$$
$$= (I - \mu A)^{t+1} x^{(0)}$$

and we care about $\|x^{(t+1)} - x^*\|$

$$\|x^{(t+1)} - x^*\| = \|x^{(t+1)} - 0\|$$
$$= \|(I - \mu A)^{t+1} x^{(0)}\|$$

$$\leq \|(I-\mu A)^{t+1}\| \, \|x^{(0)}\|$$

↳ by the fact that

$$\|Mv\| \leq \underbrace{\|M\|}_{\substack{= \text{max envalue} \\ \text{of } M}} \|v\|$$

M is PSD

$$Iv = v$$

$$-\mu A v = -\mu \lambda v$$

$$(I-\mu A)v = (1-\mu\lambda) v$$

$$= \left[\text{max eigenvalue of } I-\mu A\right]^{t+1} \|x^{(0)}\|$$

$$= \cdots = \max_i |1-\mu\lambda_i|^{t+1} \|x^{(0)}\|$$

λ_i are the eigenvalues of A.

So for this example ($f(x) = \frac{1}{2}x^T A x$)

$$\|x^{(t+1)} - x^*\| \leq \underbrace{\max_i |1-\mu\lambda_i|}_{= \max\{1-\mu\lambda_{min}, \, \mu\lambda_{max}-1\}}{}^{t+1} \|x^{(0)}\|$$

So, to get fast convergence for GD we'd like

$\max(1-\lambda_{min}\mu, \, \lambda_{max}\mu - 1)$ to be small so that when we raise it to the power $t+1$ it gets even smaller.

Turns out, the optimal choice of μ is

$$\boxed{\mu^* = \frac{2}{\lambda_{max} + \lambda_{min}}}.$$

With this choice, the corresponding rate of convergence is

$$\boxed{\frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}} = \frac{\kappa - 1}{\kappa + 1}$$

If we define $\kappa = \frac{\lambda_{max}}{\lambda_{min}}$ = condition number of A

Then the opt. conv. rate of GD can be rewritten as

$\frac{\kappa - 1}{\kappa + 1}$ which means that when κ is large the convergence is slow

$$\|x^{(t+1)} - 0\| \le \left(\frac{\kappa-1}{\kappa+1}\right)^{t+1} \|x^{(0)}\|$$

On the other hand, with **momentum**:

$$x^{(t+1)} = x^{(t)} - \mu \nabla f(x^{(t)}) + \beta\left(x^{(t)} - x^{(t-1)}\right)$$

Here, the opt. choice of μ & β gives you a convergence rate of $\sqrt{\beta}$ (like in the single variable case from last lecture)

with $\sqrt{\beta} = \dfrac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

$\Rightarrow$ convergence is accelerated compared to GD

How do we see this?

Consider for example $\kappa = 100$

then GD would have

$$\|x^{(t+1)} - x^*\| \leq \left(\frac{100-1}{100+1}\right)^{t+1} \|x^{(0)}\|$$

$$= \left(\frac{99}{101}\right)^{t+1} (\text{Junk})$$

while momentum:

$$\|x^{(t+1)} - x^*\| \leq \left(\frac{\sqrt{100}-1}{\sqrt{100}+1}\right)^{t+1} (\text{Junk})$$

$$= \left(\frac{9}{11}\right)^{t+1} (\text{Junk}$$

—x—

Variation : Nesterov's Accelaration

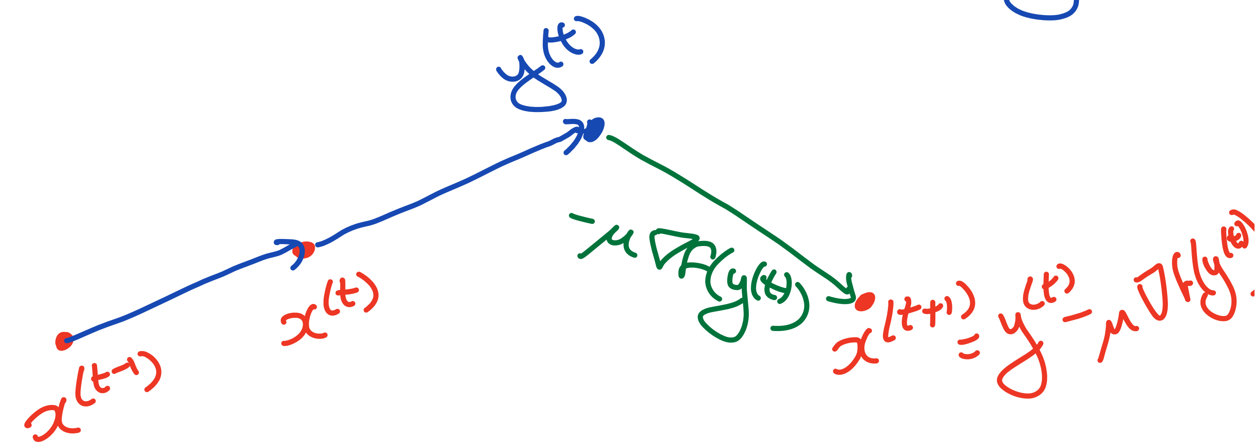
$$y^{(t)} = x^{(t)} + \beta\left(x^{(t)} - x^{(t-1)}\right) \quad \text{— } 1$$

$$x^{(t+1)} = y^{(t)} - \mu \nabla f\left(y^{(t)}\right) \quad \text{— } 2$$

Interpretation:

① Take a "momentum step" so you land at $y^{(t)}$

② Take a GD step from $y^{(t)}$.



We can of course combine ① & ② to get

$$\boxed{\begin{aligned} x^{(t+1)} = x^{(t)} + \beta\left(x^{(t)} - x^{(t-1)}\right) \\ - \mu\nabla f\left(x^{(t)} + \beta\left(x^{(t)} - x^{(t-1)}\right)\right) \end{aligned}}$$

↳ nesterov's accelaration

Converges at an accelerated rate for <u>any</u> convex problem

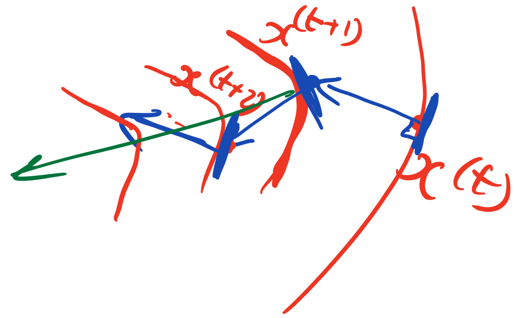
with rate $= \sqrt{\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}}$

arises with the optimal choice of

$\mu = \frac{1}{\lambda_{max}}$, $\beta = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

in the quadratic case

- used in practice
- speed up GD significantly.


$x^{(t+1)}$
$x^{(t+2)}$
$x(t)$