Optimization

First order methods: Weaknesses of Gradient Descent

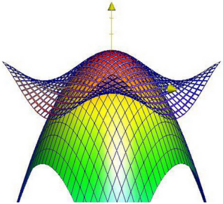
Learning goals

- LEARNING GOAL 1
- LEARNING GOAL 2

REMINDER: LOCAL QUADRATIC GEOMETRY

Every function can be locally approximated by a quadratic function via Taylor approximation:

$$f(\mathbf{x}) \approx f(\tilde{\boldsymbol{x}}) + \nabla f(\tilde{\boldsymbol{x}})^\top(\mathbf{x} - \tilde{\boldsymbol{x}}) + \frac{1}{2}(\mathbf{x} - \tilde{\boldsymbol{x}})^\top \nabla^2 f(\tilde{\boldsymbol{x}})(\mathbf{x} - \tilde{\boldsymbol{x}})$$



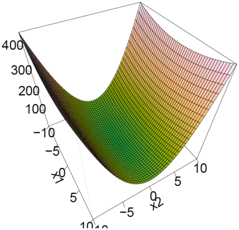
f is shown as the hollow grid and its second-order approximation at $(0, 0)$ as a continuous surface. Source: `daniloroccatano.blog`.

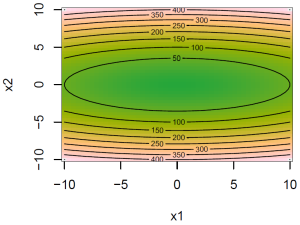
REMINDER: LOCAL QUADRATIC GEOMETRY

We will therefore look at the Hessian $\boldsymbol{H} = \nabla^2 f(\mathbf{x}^{[t]})$ at a given iteration of gradient descent and discuss weaknesses of GD depending on the local curvature of a function.

Recall:

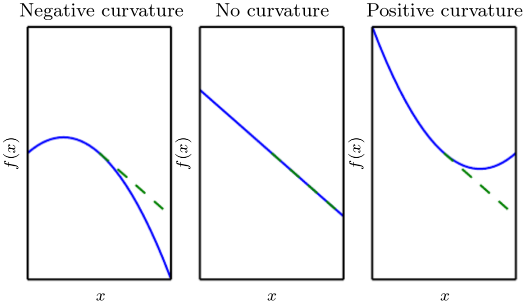
- The eigenvector $\mathbf{v}_{\text{max}}$ ($\mathbf{v}_{\text{min}}$) belonging to the largest (smallest) eigenvalue λ_{max} (λ_{min}) is the direction of max (min) curvature.
- We call the Hessian ill-conditioned if the ratio $\kappa(\mathbf{A}) = \frac{|\lambda_{\text{max}}|}{|\lambda_{\text{min}}|}$ is high.





EFFECTS OF CURVATURE

Intuitively, the curvature of a function determines the outcome of a GD step…



Source: Goodfellow *et al.*, (2016), ch. 4

Quadratic objective function $f(\mathbf{x})$ with various curvatures. The dashed line indicates the first order Taylor approximation. Left: The cost function decreases faster than the gradient predicts; Middle: The gradient predicts the decrease correctly; Right: The function decreases more slowly than expected and begins to increase.

CURVATURE AND STEP-SIZE IN GD

In the worst case, the Hessian is ill-conditioned. What does this mean for GD?

- Let us consider the second-order Taylor approximation of $f(\mathbf{x})$ around $\tilde{\boldsymbol{x}}$ (with gradient $\mathbf{g}$)

$$f(\mathbf{x}) \approx f(\tilde{\boldsymbol{x}}) + (\mathbf{x} - \tilde{\boldsymbol{x}})^\top \boldsymbol{g} + \frac{1}{2}(\mathbf{x} - \tilde{\boldsymbol{x}})^\top \boldsymbol{H}(\mathbf{x} - \tilde{\boldsymbol{x}})$$

- One GD step with a learning rate α yields new parameters $\tilde{\boldsymbol{x}} - \alpha\mathbf{g}$ and a new approximated function value

$$f(\tilde{\boldsymbol{x}} - \alpha\mathbf{g}) \approx f(\tilde{\boldsymbol{x}}) - \alpha\mathbf{g}^\top\mathbf{g} + \frac{1}{2}\alpha^2\mathbf{g}^\top\mathbf{H}\,\mathbf{g}\,.$$

- Theoretically, if $\mathbf{g}^\top\mathbf{H}\mathbf{g}$ is positive, we can solve the equation above for the optimal step size which corresponds to

$$\alpha^* = \frac{\mathbf{g}^\top\mathbf{g}}{\mathbf{g}^\top\mathbf{H}\,\mathbf{g}}\,.$$

CURVATURE AND STEP-SIZE IN GD

- Let us assume the gradient $\mathbf{g}$ points into the direction of $\boldsymbol{v}_{\text{max}}$ (i.e. the direction of highest curvature), the optimal step size is given by

$$\alpha^* = \frac{\mathbf{g}^\top \mathbf{g}}{\mathbf{g}^\top \mathbf{H}\,\mathbf{g}} = \frac{\mathbf{g}^\top \mathbf{g}}{\lambda_{\text{max}} \mathbf{g}^\top \mathbf{g}} = \frac{1}{\lambda_{\text{max}}},$$

  which is very small. Choosing a too large step-size is bad, as it will make us “overshoot” the stationary point.

- If, on the other hand, $\mathbf{g}$ points into the direction of the lowest curvature, the optimal step size is

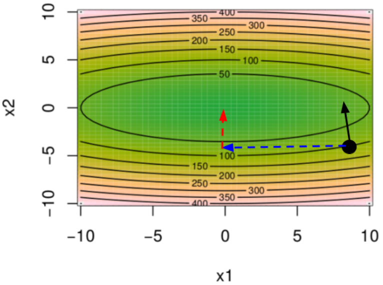
$$\alpha^* = \frac{1}{\lambda_{\text{min}}},$$

  which corresponds to the largest possible optimal step-size.

- We summarize: We want to perform big steps in directions of low curvature, but small steps in directions of high curvature.

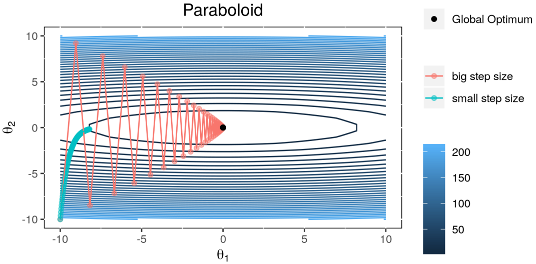
CURVATURE AND STEP-SIZE IN GD

- But what if the gradient does not point into the direction of one of the eigenvectors?
- Let us consider the 2-dimensional case: We can decompose the direction of $\boldsymbol{g}$ (black) into the two eigenvectors $\boldsymbol{v}_{\text{max}}$ and $\boldsymbol{v}_{\text{min}}$
- It would be optimal to perform a **big** step into the direction of the smallest curvature $\boldsymbol{v}_{\text{min}}$, but a **small** step into the direction of $\boldsymbol{v}_{\text{max}}$, but the gradient points into a completely different direction.



CURVATURE AND STEP-SIZE IN GD

- GD is unaware of large differences in curvature, and can only walk into the direction of the gradient.
- Choosing a too large step-size will then cause the descent direction change frequently (“jumping around”).
- α needs to be small enough, which results in a low progress.



The contour lines show a quadratic risk function with a poorly conditioned Hessian matrix. The plot shows the progress of gradient descent with a small step-size vs. larger step-size. In both cases, convergence to the global optimum is rather slow.

CURVATURE AND STEP-SIZE IN GD

- In the worst case, ill-conditioning of the Hessian matrix and a too big step-size will cause the risk to increase

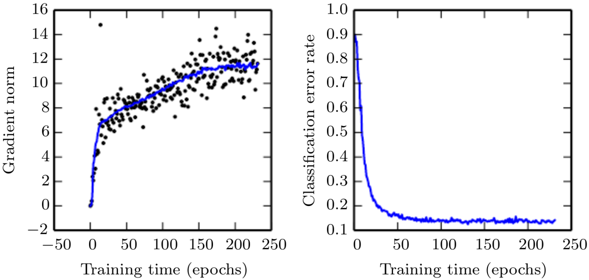
$$f(\tilde{\boldsymbol{x}} - \alpha \mathbf{g}) \approx f(\tilde{\boldsymbol{x}}) - \alpha \mathbf{g}^\top \mathbf{g} + \frac{1}{2}\alpha^2 \mathbf{g}^\top \mathbf{H}\, \mathbf{g}\,.$$

  which happens if

$$\frac{1}{2}\alpha^2 \mathbf{g}^\top \mathbf{H}\, \mathbf{g} > \alpha \mathbf{g}^\top \mathbf{g}.$$

- To determine whether ill-conditioning is detrimental to the training, the squared gradient norm $\mathbf{g}^\top \mathbf{g}$ and the risk can be monitored.

CURVATURE AND STEP-SIZE IN GD



Source: Goodfellow, ch. 6

- Gradient norms **increase** over time, showing that the training process is not converging to a stationary point $\mathbf{g} = 0$.
- At the same time, we observe that the risk is approx. constant, but the gradient norm increases

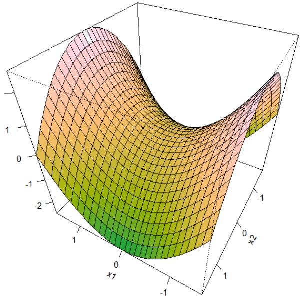
$$\underbrace{f(\tilde{\boldsymbol{x}} - \alpha \mathbf{g})}_{\text{approx. constant}} \approx f(\tilde{\boldsymbol{x}}) - \underbrace{\alpha \mathbf{g}^\top \mathbf{g}}_{\text{increase}} + \frac{1}{2} \underbrace{\alpha^2 \mathbf{g}^\top \mathbf{H} \, \mathbf{g}}_{\to \text{increase}} .$$

GD AT SADDLE POINTS

Example:

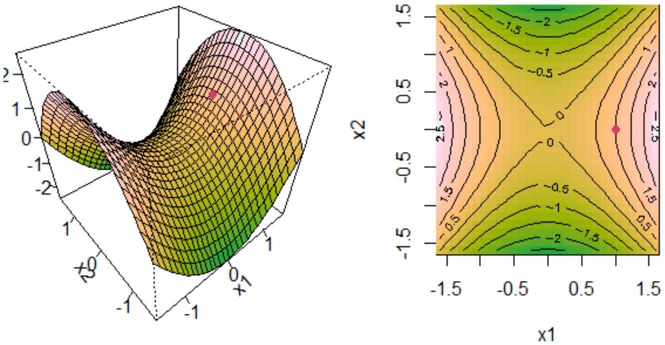
$$f(x_1, x_2) = x_1^2 - x_2^2$$

Along x_1, the function curves upwards (eigenvector of the Hessian with positive eigenvalue). Along x_2, the function curves downwards (eigenvector of the Hessian with negative eigenvalue).



EXAMPLE: SADDLE POINT WITH GD

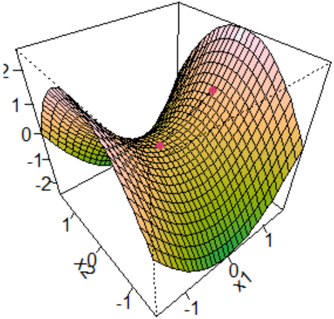
- So how do saddle points impair optimization?
- First-order algorithms that use only gradient information **might** get stuck in saddle points.

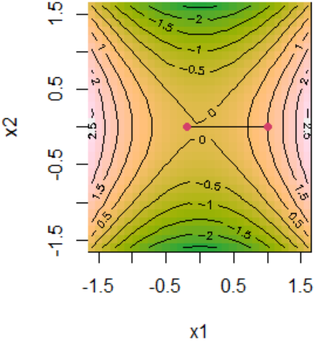


Red dot: Starting location

EXAMPLE: SADDLE POINT WITH GD

- So how do saddle points impair optimization?
- First-order algorithms that use only gradient information **might** get stuck in saddle points.

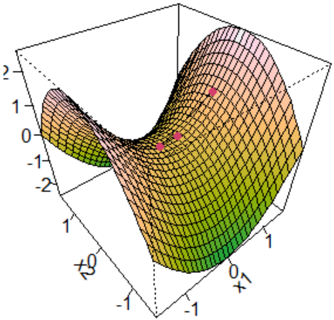


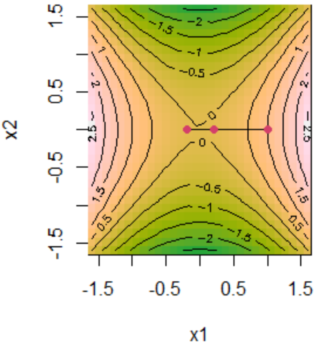


First step...

EXAMPLE: SADDLE POINT WITH GD

- So how do saddle points impair optimization?
- First-order algorithms that use only gradient information **might** get stuck in saddle points.

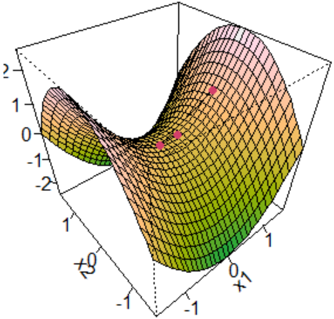


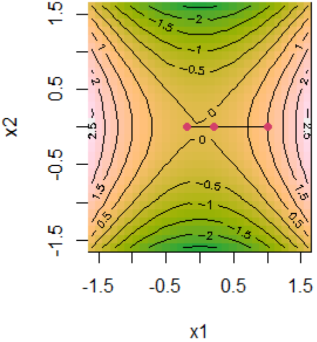


...second step...

EXAMPLE: SADDLE POINT WITH GD

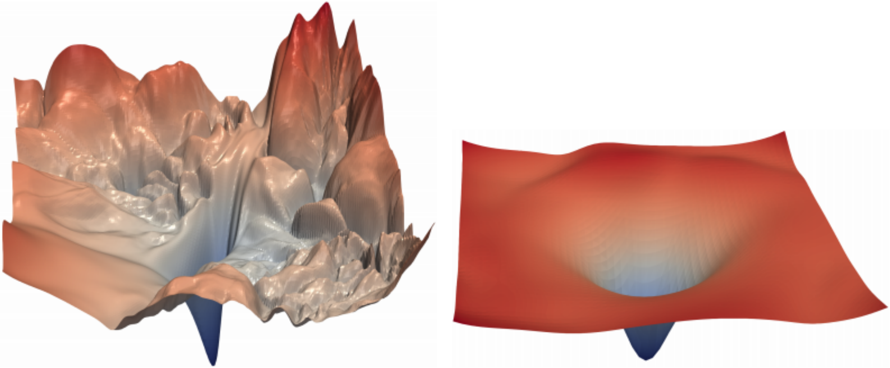
- So how do saddle points impair optimization?
- First-order algorithms that use only gradient information **might** get stuck in saddle points.





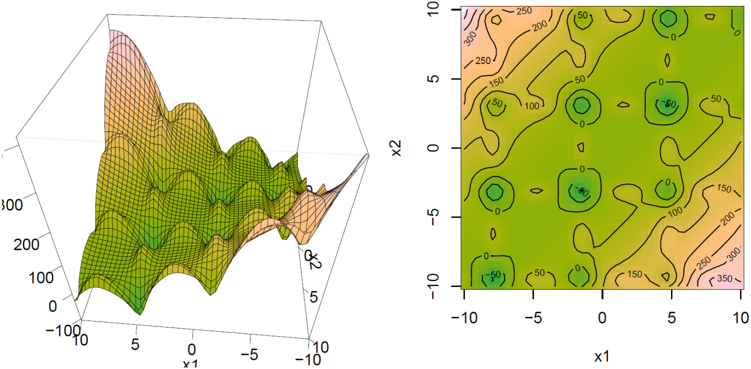
...tenth step got stuck and cannot escape the saddle point!

UNIMODAL VS. MULTIMODAL LOSS SURFACES



Left: Multimodal loss surface with saddle points; Right: (Nearly) unimodal loss surface
(Hao Li et al. (2017))

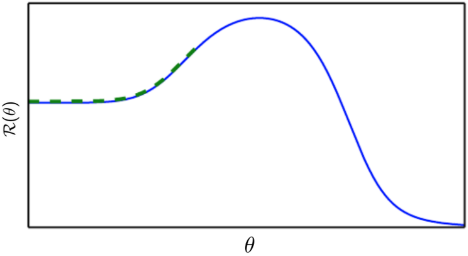
UNIMODAL VS. MULTIMODAL LOSS SURFACES



Potential snippet from a loss surface with many local minima

ONLY LOCALLY OPTIMAL MOVES

- If the training algorithm makes only *locally* optimal moves (as in gradient descent), it may move away from regions of *much* lower cost.

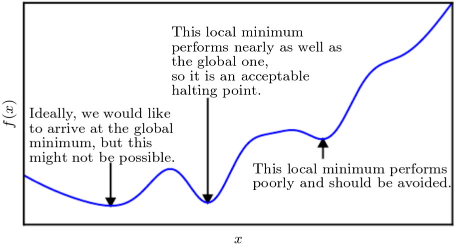


Source: Goodfellow, Ch. 8

- In the figure above, initializing the parameter on the “wrong” side of the hill will result in suboptimal performance.
- In higher dimensions, however, it may be possible for gradient descent to go around the hill but such a trajectory might be very long and result in excessive training time.

LOCAL MINIMA

- In practice only local minima with a high value compared to the global minimium are problematic.



Source: Goodfellow, Ch. 4

- In DL, literature suspects that most local minima have low empirical risk. (Y. Dauphin et al. (2014))
- Simple test: Norm of gradient should get close to zero.