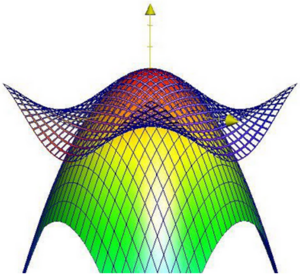
Optimization in Machine Learning

First order methods: Weaknesses of GD – Curvature



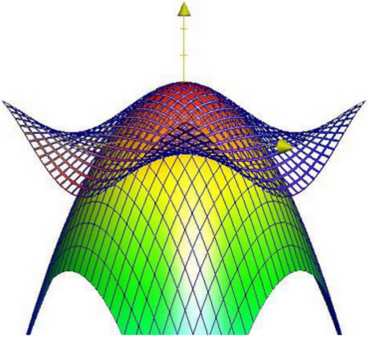
Learning goals

- Effects of curvature
- Step size effect in GD

REMINDER: LOCAL QUADRATIC GEOMETRY

Approx smooth function locally via 2nd order Taylor:

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



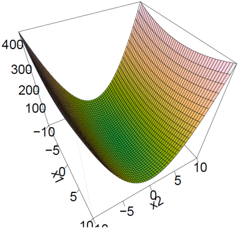
Source: `daniloroccatano.blog`.

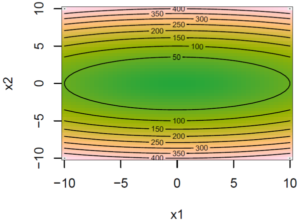
REMINDER: LOCAL QUADRATIC GEOMETRY

Study Hessian $\boldsymbol{H} = \nabla^2 f(\mathbf{x}^{[t]})$ in GD to discuss effect of curvature

Recall:

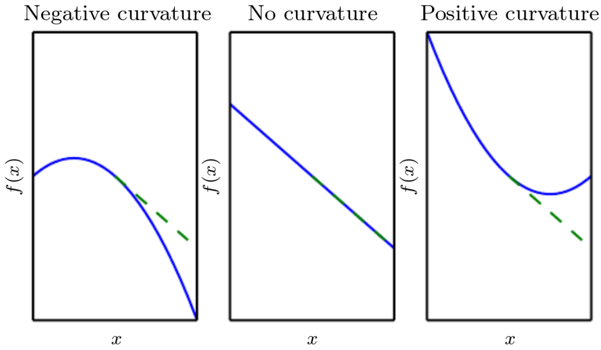
- E-vec $\mathbf{v}_{\max}$ ($\mathbf{v}_{\min}$) for E-val $\lambda_{\max}$ ($\lambda_{\min}$) is dir. of max (min) curvature
- $\boldsymbol{H}$ called ill-conditioned if ratio $\kappa(\mathbf{A}) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$ is high





EFFECTS OF CURVATURE

Intuitively, curvature of function determines outcome of GD step...



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

Quadratic objective $f(\mathbf{x})$. Dashed line = 1st order Taylor. Left: f decreases faster than grad predicts; Middle: grad predicts decrease correctly; Right: f decreases more slowly than grad predicts, then increases.

CURVATURE AND STEP-SIZE IN GD

Worst case: $\boldsymbol{H}$ is ill-conditioned. What does this mean for GD?

- 2nd order Taylor of $f(\mathbf{x})$ around $\tilde{\boldsymbol{x}}$ (with grad $\mathbf{g}$)

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

- One GD step with step size α yields new parameters $\tilde{\boldsymbol{x}} - \alpha\mathbf{g}$ and new approx function value

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

- If $\mathbf{g}^\top\boldsymbol{H}\mathbf{g} > 0$, we can solve above for optimal step size α:

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

CURVATURE AND STEP-SIZE IN GD

- Let's assume grad $\mathbf{g}$ points into dir. of $\boldsymbol{v}_{\text{max}}$ (so highest curvature), then optimal step size is:

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

  which is small. Large α will make us "overshoot".

- OTOH: if $\mathbf{g}$ points into dir. of smallest curvature:

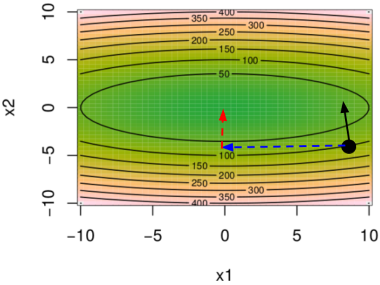
$$\alpha^* = \frac{1}{\lambda_{\text{min}}},$$

  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.

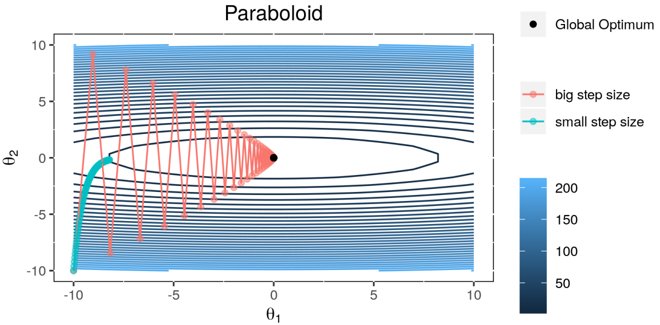
CURVATURE AND STEP-SIZE IN GD

- But what if g doesn't align with these E-vecs?
- Let us consider the 2-dimensional case: We can decompose the direction of $\boldsymbol{g}$ (black) into the two eigenvectors $\boldsymbol{v}_{\max}$ and $\boldsymbol{v}_{\min}$
- It would be optimal to perform a **big** step into the direction of the smallest curvature $\boldsymbol{v}_{\min}$, but a **small** step into the direction of $\boldsymbol{v}_{\max}$, 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 slow progress.



Contour lines = poorly cond. quadratic f. GD with small vs. large α. For both, convergence to optimum is 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 objective to increase

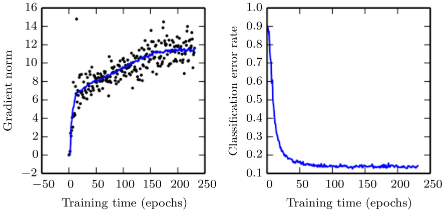
$$f(\tilde{\boldsymbol{x}} - \alpha \mathbf{g}) \approx f(\tilde{\boldsymbol{x}}) - \alpha \mathbf{g}^\top \mathbf{g} + \frac{1}{2}\alpha^2 \mathbf{g}^\top \boldsymbol{H} \mathbf{g}\,.$$

which happens if

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

CURVATURE AND STEP-SIZE IN GD

- To see if ill-conditioning is problematic in the opt run, we can monitor squared gradient norm $\mathbf{g}^\top \mathbf{g}$ and f.



Source: Goodfellow, ch. 6

- Gradient norms **increase** over time, showing that the training process is not converging to stationary $\mathbf{g} = 0$.
- But f (risk) approx. constant.

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