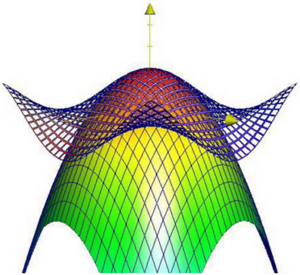
Optimization in Machine Learning

First order methods
Weaknesses of GD – Curvature



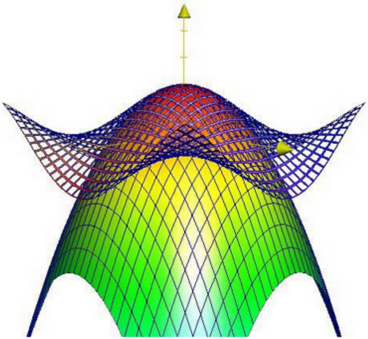
Learning goals

- Effects of curvature
- Step size effect in GD

REMINDER: LOCAL QUADRATIC GEOMETRY

- Locally approximate smooth function by quadratic Taylor polynomial:

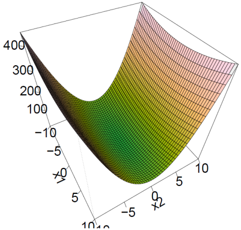
$$f(\mathbf{x}) \approx f(\tilde{\mathbf{x}}) + \nabla f(\tilde{\mathbf{x}})^\top (\mathbf{x} - \tilde{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \tilde{\mathbf{x}})^\top \nabla^2 f(\tilde{\mathbf{x}})(\mathbf{x} - \tilde{\mathbf{x}})$$

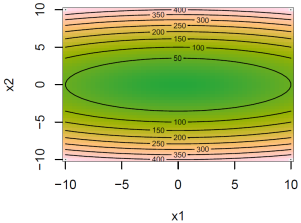


Source: `daniloroccatano.blog`.

REMINDER: LOCAL QUADRATIC GEOMETRY

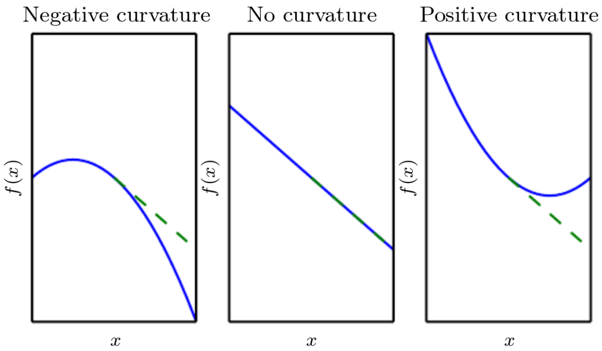
- Study Hessian $\mathbf{H} = \nabla^2 f(\mathbf{x}^{[t]})$ in GD to discuss effect of curvature
- **Recall** for quadratic forms:
  - Eigenvector $\mathbf{v}_{\text{max}}$ ($\mathbf{v}_{\text{min}}$) is direction of largest (smallest) curvature
  - $\mathbf{H}$ called ill-conditioned if $\kappa(\mathbf{H}) = |\lambda_{\text{max}}|/|\lambda_{\text{min}}|$ is large





EFFECTS OF CURVATURE

- Intuitively, curvature determines reliability of a GD step



Quadratic objective f (blue) with gradient approximation (dashed green). **Left:** f decreases faster than ∇f predicts. **Center:** ∇f predicts decrease correctly. **Right:** f decreases more slowly than ∇f predicts. (Source: Goodfellow et al., 2016)

CURVATURE AND STEP SIZE IN GD

- **Worst case: H** is ill-conditioned.What does this mean for GD?
  - Quadratic Taylor polynomial of f around $\tilde{\mathbf{x}}$ (with gradient $\mathbf{g} = \nabla f$)

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

  - GD step with step size $\alpha > 0$ yields

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

  - If $\mathbf{g}^\top \boldsymbol{H}\mathbf{g} > 0$, we can solve 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

- If **g** points along $\mathbf{v}_{\text{max}}$ (largest curvature), optimal step size is

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

  $\Rightarrow$ *Large* step sizes can be problematic.

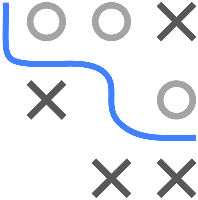
- If **g** points along $\mathbf{v}_{\text{min}}$ (smallest curvature), then analogously

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

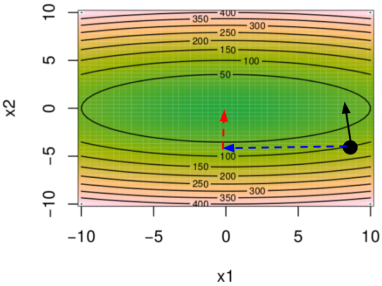
  $\Rightarrow$ *Small* step sizes can be problematic.

- **Ideally**: Perform large step along $\mathbf{v}_{\text{min}}$ but small step along $\mathbf{v}_{\text{max}}$.

CURVATURE AND STEP SIZE IN GD



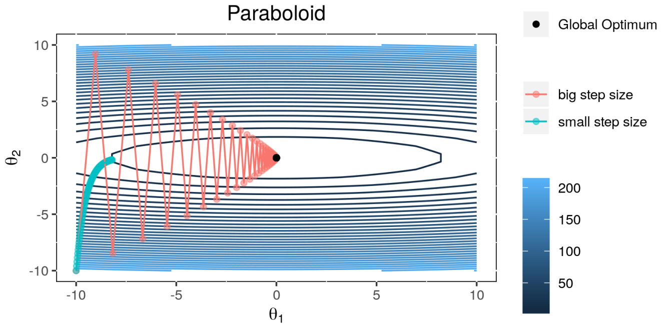
- What if **g** is not aligned with eigenvectors?
- Consider 2D case: Decompose **g** (black) into $\color{red}{\mathbf{v}_{\text{max}}}$ and $\color{blue}{\mathbf{v}_{\text{min}}}$



- Ideally, perform **large** step along $\color{blue}{\mathbf{v}_{\text{min}}}$ but **small** step along $\color{red}{\mathbf{v}_{\text{max}}}$
- However, gradient almost only points along $\color{red}{\mathbf{v}_{\text{max}}}$

CURVATURE AND STEP SIZE IN GD

- GD is not aware of curvatures and can only walk along **g**
- Large step sizes result in “zig-zag” behaviour.
- Small step sizes result in weak performance.



Poorly conditioned quadratic form. GD with large (red) and small (blue) step size.
For both, convergence to optimum is slow.

CURVATURE AND STEP SIZE IN GD

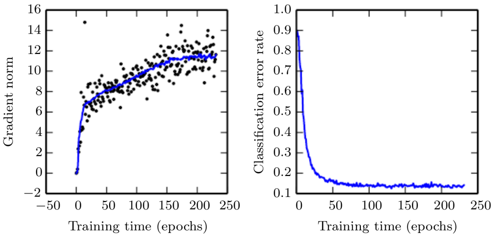
- Large step sizes for ill-conditioned Hessian can even increase

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

  if

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

- Ill-conditioning in practice: Monitor gradient norm and objective



Source: Goodfellow et al., 2016

CURVATURE AND STEP SIZE IN GD

- If gradient norms $\|\mathbf{g}\|$ increase, GD is not converging since $\mathbf{g} \neq 0$.
- Even if $\|\mathbf{g}\|$ increases, objective may stay approximately constant:

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