Lecture 21

- Practical gradient computations
- Finite difference approximations
- Matrix calculus
- Automatic differentiation

Project report due dates

- These are due on Monday @ 5pm like normal 1

Quiz

- Is there a function you'd like to learn how to take the gradient of?
- Is there a specific topic in this area you'd like me to cover?

- Kronecker Product or trace
- Matrix Polynomials
- Chebyshev Polynomials (Rosenbrock)
- Very Complex problem
- Argmin / Arg max
- $Ax^3 + x^{Ax}$
- ~~Auto-diff + Autodiff~~ Fast!

- Approx. Algorith
- ✓ Gradient Approxim
- ✓ finite diff

- Softmax?

- Non-diff. function

- MPC w/ $x^T Q x$

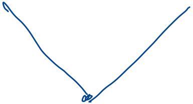
$$P(x) = \sum \frac{A_j(x)}{C(1+r)^j} - B x^b$$

- Auto-grad: linear op?

- Neural Networks

- Step functions

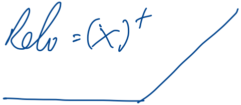
Discontinuity



$$f(x) =$$

$$\frac{\partial |x|}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \end{cases}$$

$$\text{Relu} = (x)^+$$



$$\frac{\partial \text{Relu}}{\partial x} = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{else.} \end{cases}$$

Matrix Calculus

- it's boring
- it's detail heavy
- it's Easy to make Mistakes!

true for all gradient Computations!

How do you know if you got your gradient correct?

① finite difference approximations are super-easy to compute!

Recall

$$f'(\alpha) \approx \frac{f(\alpha + h) - f(\alpha)}{h}$$

What is h? Use $\sqrt{\varepsilon} = \sqrt{2^{-52}}$ for 64-bit float.

$f(x) : \mathbb{R}^n \to \mathbb{R}$

$$\frac{\partial f}{\partial x} = g(x) \approx$$

Requires $n+1$

$$\begin{bmatrix} \dfrac{f(\underline{x}+h\underline{e}_1)-f(\underline{x})}{h} \\ \dfrac{f(\underline{x}+h\underline{e}_2)-f(\underline{x})}{h} \\ \vdots \\ \dfrac{f(\underline{x}+h\underline{e}_n)-f(\underline{x})}{h} \end{bmatrix}$$

f-Evals ☹

Let $\hat{g}(x) \approx g(x)$ via finite differences

if you want to test $g(x)$ then compute

$$\|\hat{g}(x) - g(x)\|_\infty$$

and check if its small.

What $\underline{x}$? Use a random point that's feasible / your starting point, etc.

Can you go

$g(x)^T p$

Faster?

$$\approx \frac{f(x+hp) - f(x)}{h}$$

Example

$$f_3(x_1, x_2, x_3) = (x_1 x_2 x_3)^3$$

or.

$$\boxed{f(x) = g\Big(\prod_{i=1}^{n} (x_i)\Big)}$$

$$\rightarrow \frac{\partial f}{\partial x} = g'\Big(\prod (x_i)\Big)$$

$$\frac{\partial f_3}{\partial x_1} = 3(x_1 x_2 x_3)^2 \, x_2 x_3$$

$\vdots$

$$\frac{\partial f}{\partial x} = g'\Big(\prod (x_i)\Big) \cdot \begin{bmatrix} x_2 \cdots x_n \\ x_1 x_3 - x_n \\ \vdots \\ x_1 \cdots x_{n-1} \end{bmatrix}$$

$$\prod_{i=1}^{n} (x_i) / x_i$$

$\Pi(x) X^{-1} \underline{e}$

$$f(x) = \sum_{i=1}^{n} \log(x_i)$$

$$\frac{\partial f}{\partial x} = X^{-1} \underline{e}$$

why $\frac{\partial f}{\partial x_i} = \frac{1}{x_i}$

- Rules needed for $e^{\log(x)^T A \log(x)}$

$$\begin{aligned}
d\mathbf{A} &= 0 \quad (\text{for constant } \mathbf{A}) \\
d(\alpha\mathbf{X}) &= \alpha d\mathbf{X} \\
d(\mathbf{X}+\mathbf{Y}) &= d\mathbf{X} + d\mathbf{Y} \\
d(\mathrm{tr}(\mathbf{X})) &= \mathrm{tr}(d\mathbf{X}) \\
d(\mathbf{X}\mathbf{Y}) &= (d\mathbf{X})\mathbf{Y} + \mathbf{X}d\mathbf{Y} \\
d(\mathbf{X}\otimes\mathbf{Y}) &= (d\mathbf{X})\otimes\mathbf{Y} + \mathbf{X}\otimes d\mathbf{Y} \quad (\text{see section} \\
d(\mathbf{X}\circ\mathbf{Y}) &= (d\mathbf{X})\circ\mathbf{Y} + \mathbf{X}\circ d\mathbf{Y} \quad (\text{see section} \\
d\mathbf{X}^{-1} &= -\mathbf{X}^{-1}(d\mathbf{X})\mathbf{X}^{-1} \\
\boxed{d|\mathbf{X}| = |\mathbf{X}|\,\mathrm{tr}(\mathbf{X}^{-1}d\mathbf{X})} \\
\cancel{d\log|\mathbf{X}| = \mathrm{tr}(\mathbf{X}^{-1}d\mathbf{X})} \\
d\mathbf{X}^{\star} &= (d\mathbf{X})^{\star}
\end{aligned}$$

$dy = a\,dx$	$d\mathbf{y} = \mathbf{a}\,dx$	$d\mathbf{Y} = \mathbf{A}\,dx$
$dy = \mathbf{a}\,d\mathbf{x}$	$d\mathbf{y} = \mathbf{A}\,d\mathbf{x}$	
$dy = \mathrm{tr}(\mathbf{A}\,d\mathbf{X})$		

n 2)

5)

$f(X) = \log\det(X))$

$\frac{df}{\partial X} \Rightarrow$

$df = \frac{1}{\det(X)} \cdot d\det(X)$

$= \cancel{\frac{1}{\det(X)} \cdot \det(X)} \cdot tr(X^{-1} dX)$

$= tr(X^{-1} \cdot dX)$

$\Rightarrow \frac{df}{\partial X} = X^{-1}$

Matrix Calc Rules will
often have a matrix in
product!

$d\,|F(X)| = d(\det$

$= |F(X)| \cdot F$

$A \bullet B =$ Matrix inner-

$A \bullet B = \sum_{ij} A_{ij} B_{ij}$

nar--

$(F(x)))$

$(x)^{-1} \cdot dF$

product

$= \operatorname{tr}(A^T B)$

$= \operatorname{vec}(A)^T \operatorname{vec}(B)$

$f(A + dA) =$

$f(A) + g(A) \cdot dA + \ldots$

$g(A) \cdot dA$ vs.

$dA \cdot g(A)$

Somethin like a graph

$f(W_1, W_2) =$

$W_2 \, \big(C \, W_1 \, x$

$\frac{\partial f}{\partial W_2} = (W_1 x$

$\frac{\partial f}{\partial W_1} = W_2 \cdot d($

$= W_2 \cdot$

Conv. Networks.

$$\boxed{\;)^{+}} =$$

$$)^{+}$$

$$(W_1 x)^{+}$$

Suppose I have

$$f(W_1) = W_1 x = \begin{bmatrix} [W_1]_{11} x_1 + \cdots [W_1]_{1n} x_n \\ [W_1]_{21} x_1 + \cdots [W_1]_{2n} x_n \\ \vdots \end{bmatrix}$$