Georgia Tech

Machine Learning CS 4641

Optimization

Nimisha Roy

Lecturer, SCI

Director of Online Undergraduate Initiatives

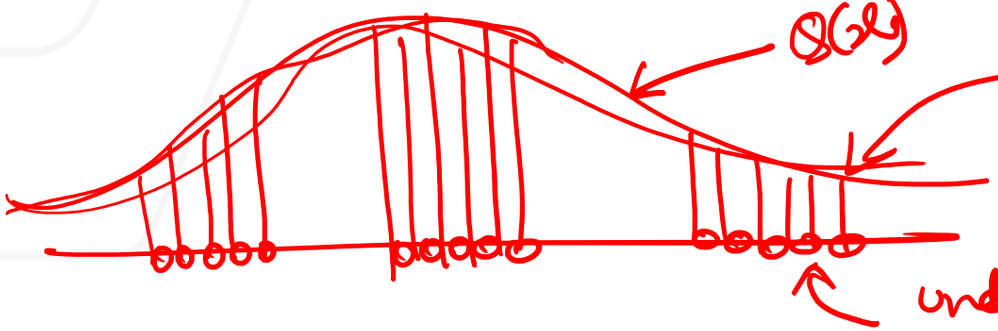
College of Computing

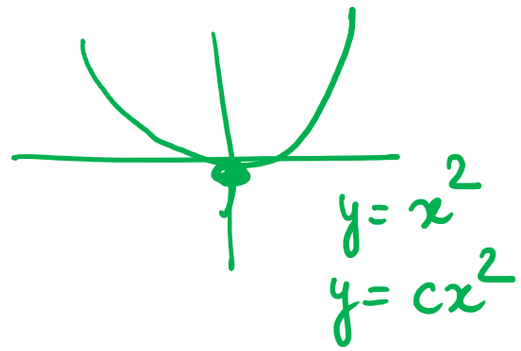
Georgia Tech

Slides adapted from Max Roozbahani, Rod Borela

$$CE = \sum_x p(x) \log \frac{1}{q(x)}$$

minimizing CE is the same as minimizing −ve log likelihood.





$$L(\theta|x) = \prod_i f(x^{(i)}|\theta)$$

$$\log L(\theta|x) = l(\theta|x) = \max \sum_i \log f(x^{(i)}|\theta)$$

$$= \min -\sum_i \log f(x^{(i)}|\theta)$$

$$= \min \frac{-1}{n} \sum_i \log f(x^{(i)}|\theta)$$

$$= \min -E[\log f(x|\theta)]$$

$$= \min -E[\log Q(x)]$$

Georgia Tech

minimizing −ve log likelihood = $\min -E[\log Q(x)]$

$$\min \; -E\left[\log P(x) - \log P(x) + \log Q(x)\right]$$

$$\min \; -E[\log P(x)] - E\left[\log \frac{Q(x)}{P(x)}\right]$$

$$\min \left(-\sum_x p(x) \log P(x) \right) \left(-\sum_x p(x) \log \frac{Q(x)}{P(x)} \right)$$

$$\min \left(H(P) + KL[P](Q) \right)$$

$$\min CE$$

$$\min. \text{ -log likelihood} = \min. CE$$

$E[a+b+c]$
$= E[a] + E[b] + E[c]$

$E[g(x)]$
$= \sum_x p(x) g(x)$

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x)\ \log q(x) = H(P) + KL[P][Q]$$

This is because:

$$H(p,q) = \mathrm{E}_p[l_i] = \mathrm{E}_p\left[\log \frac{1}{q(x_i)}\right]$$

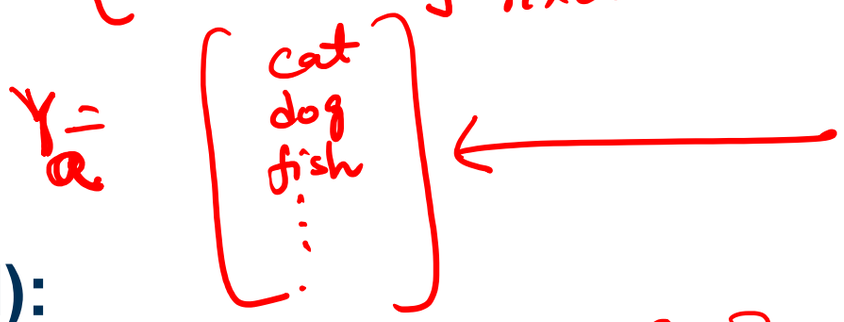
$$H(p,q) = \sum_{x_i} p(x_i)\ \log \frac{1}{q(x_i)}$$

$$H(p,q) = -\sum_{x} p(x)\ \log q(x).$$

Labeling target values: Label encoding (ordinal) and One-hot encoding

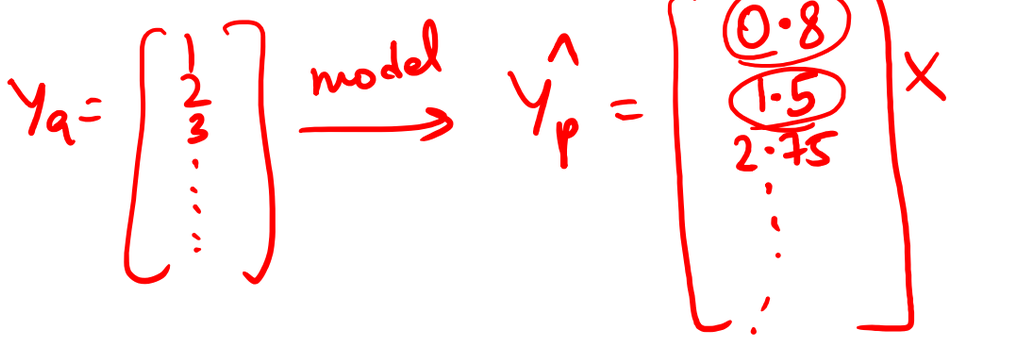
- **Input Matrix (X):**

- **Target Labels (Y):**



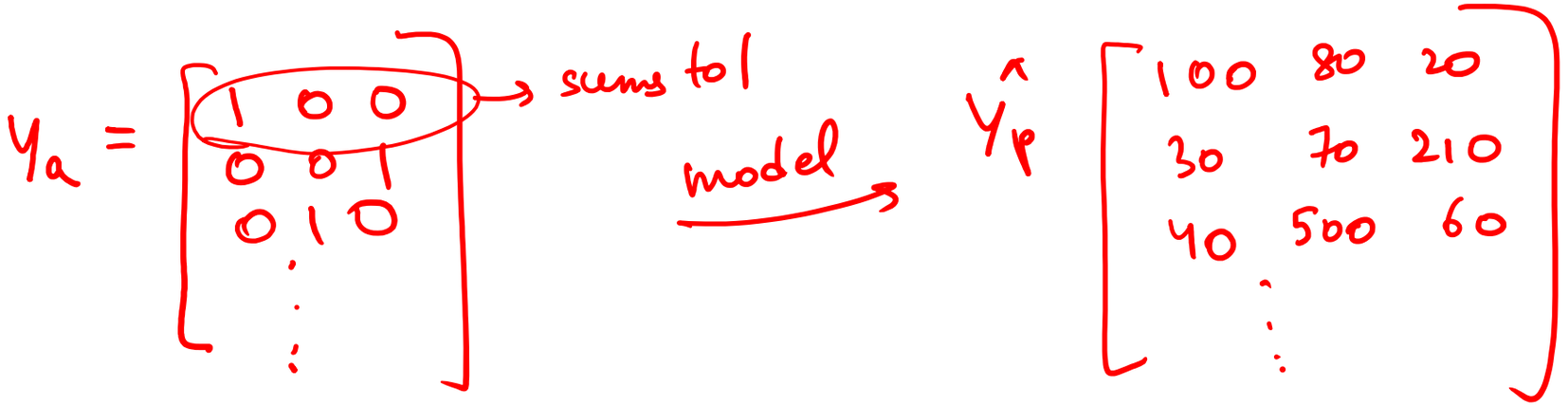
- **Label Encoding (Ordinal):**

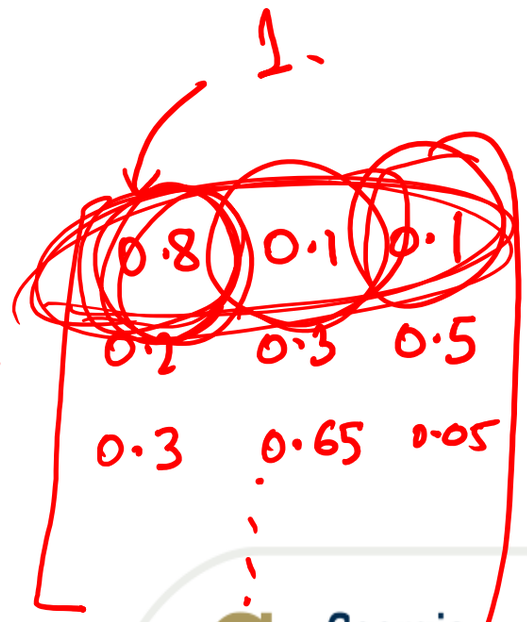
cat→1,fish→2,dog→3



- **One-Hot Encoding:**

Cat → [1 0 0], fish → [0 1 0], dog → [0 0 1]





Georgia Tech

Label encoding (One-hot encoding) Loss Computation

Minimize Objective Function:

$$\min_i \sum \| y_a^{[i]} - \hat{y}_p^{[i]} \|_2^2$$

MSE

0

$y_a = [1\ 0\ 0]$

$\hat{y_p} = [0.8\ \ 0.1\ \ 0.1]$

$(1-0.8)^2 + (0-0.1)^2 + (0-0.1)^2$

In an ideal world,

$\hat{y_p} = [1\ 0\ 0]$

error = 0 + 0 + 0 --- 0

Maximize Objective Function:

$$\max \sum_i y_a^{[i]} \cdot \hat{y}_p^{[i]T}$$

$$\min - \sum_i y_a^{[i]} \cdot \hat{y}_p^{[i]T}$$

$1 \times 0.8 + 0 \times 0.1 + 0 \times 0.1$

$= 0.8$

In an ideal world

$1 \times 1 + 0 \times 0 + 0 \times 0 = 1$

$1 + 1 + \ldots\ldots\ldots = n$

Interpretation:

- Minimum squared Error works but isn't ideal for classification.
- **Minimize negative sum of dot product is also good but Cross-Entropy / NLL** is better.

Georgia Tech

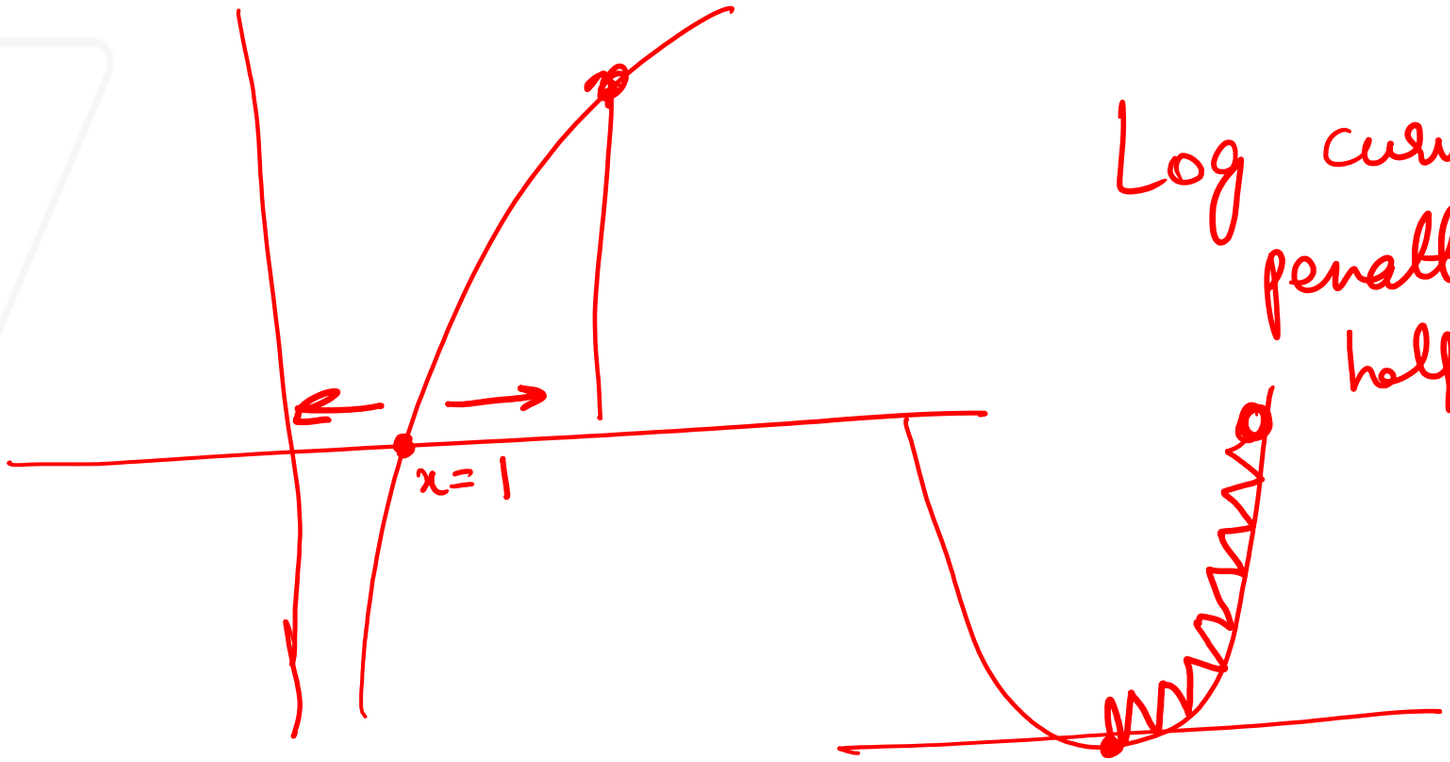
Why Cross Entropy is better than dot product

$$H(p,q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x)$$

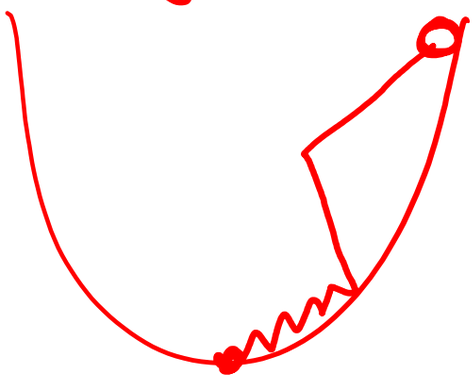
y_a $\hat{y}_p$

$$= \min_{\hat{y}_p} -\sum_i y_a^{\{i\}} \cdot \log \hat{y}_p^{\{i\}^T} \quad \text{— Case ①}$$

$$\text{Dot product} = \min_{\hat{y}_p} -\sum_i y_a^{\{i\}} \cdot \hat{y}_p^{\{i\}^T} \quad \text{— Case ②}$$



Log curve imposes a lot of penalty during optimizes which helps in easy convergence.



Logarithmic penalty encourages big update leading to faster convergence

Georgia Tech

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\mathbf{KL}[P(S)\|Q(S)] = \sum_s P(s)\log\frac{P(s)}{Q(s)}$$

$$= \underbrace{\sum_s P(s)\log\frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P] = H(P,Q) - H(P)$$

KL Divergence is a **KIND OF** distance measurement

Excess cost in bits paid by encoding according to Q instead of P.

$$-\mathbf{KL}[P\|Q] = \sum_s P(s)\log\frac{Q(s)}{P(s)}$$

log function is concave or convex?

$$\sum_s P(s)\log\frac{Q(s)}{P(s)} \leq \log\sum_s P(s)\frac{Q(s)}{P(s)}$$

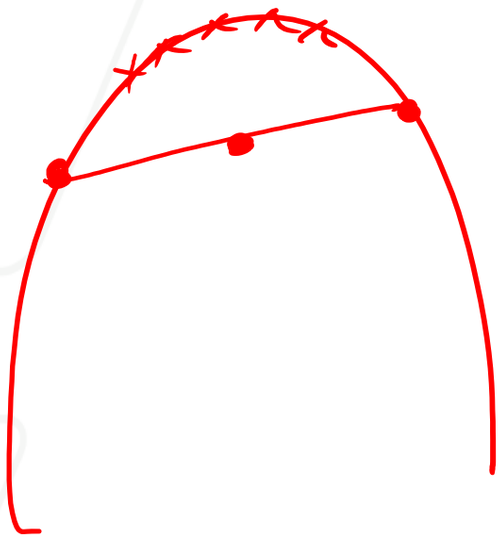
By Jensen Inequality

$$= \log\sum_s Q(s) = \log 1 = 0$$

So $\mathbf{KL}[P\|Q] \geq 0$. Equality iff $P = Q$

When $P = Q$, $KL[P||Q] = 0$

Concave Function: Jensen Inequality



$$\mathbb{E}[f(X)] \leq f(\mathbb{E}[X])$$

$f(x) = -x^2$

$f'(x) = -2x$

$f''(x) = -2 \quad < 0 \rightarrow$ concave

Strictly concave functions have a unique global maximum if it exists

Convex Function: Jensen Inequality

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

$f''(x) > 0$

$f(x) = x^2$

$x = [1 \quad 2]$

$p(x) = [\frac{1}{2} \quad \frac{1}{2}]$

$$E[X] = \sum_x x \cdot p(x) = 1 \times \frac{1}{2} + 2 \times \frac{1}{2} = 1.5$$

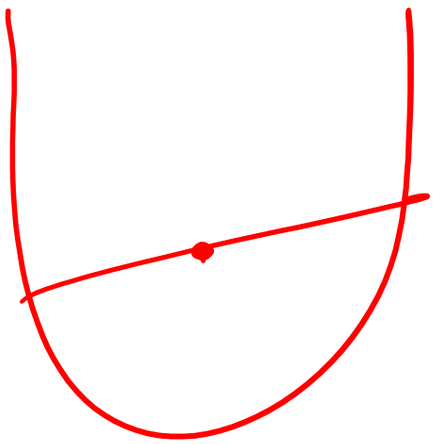
$$E[f(x)] = \sum_x f(x) p(x) = 1^2 \times \frac{1}{2} + 2^2 \times \frac{1}{2} = 2.5$$

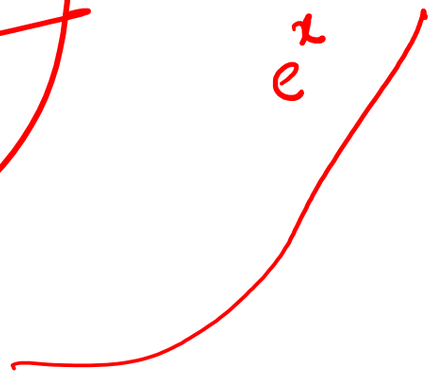
$$f(E[x]) = 1.5^2 = 2.25$$

Strictly convex functions have unique global minimum if it exists

$f(x) = 3x + 4y$

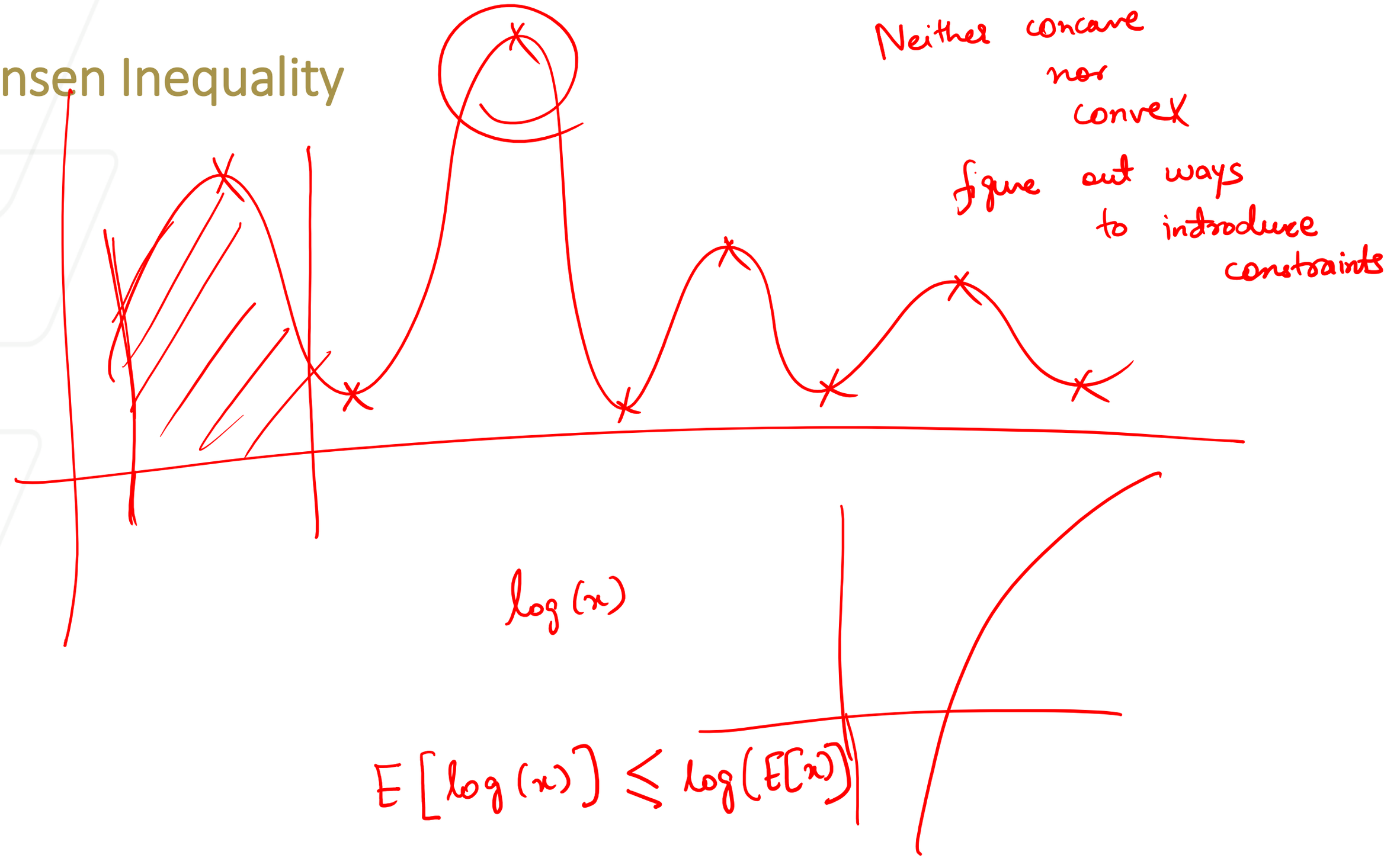
Hessian Matrix $= \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$





det(Hessian)
> 0 Convex
< 0 Concave

Jensen Inequality



KL Divergence is always non negative

$$-KL[P][Q] = \sum_x p(x) \log \frac{q(x)}{p(x)}$$

$g(x)$

$$= \leq \sum_x p(x) \log g(x) = E[\log g(x)]$$

$$\leq \log(E[g(x)])$$

$$\leq \log\left(E\left[\frac{q(x)}{p(x)}\right]\right)$$

$$\leq \log\left(\sum_x p(x) \cdot \frac{q(x)}{p(x)}\right)$$

$$\leq \log \sum_x q(x)$$

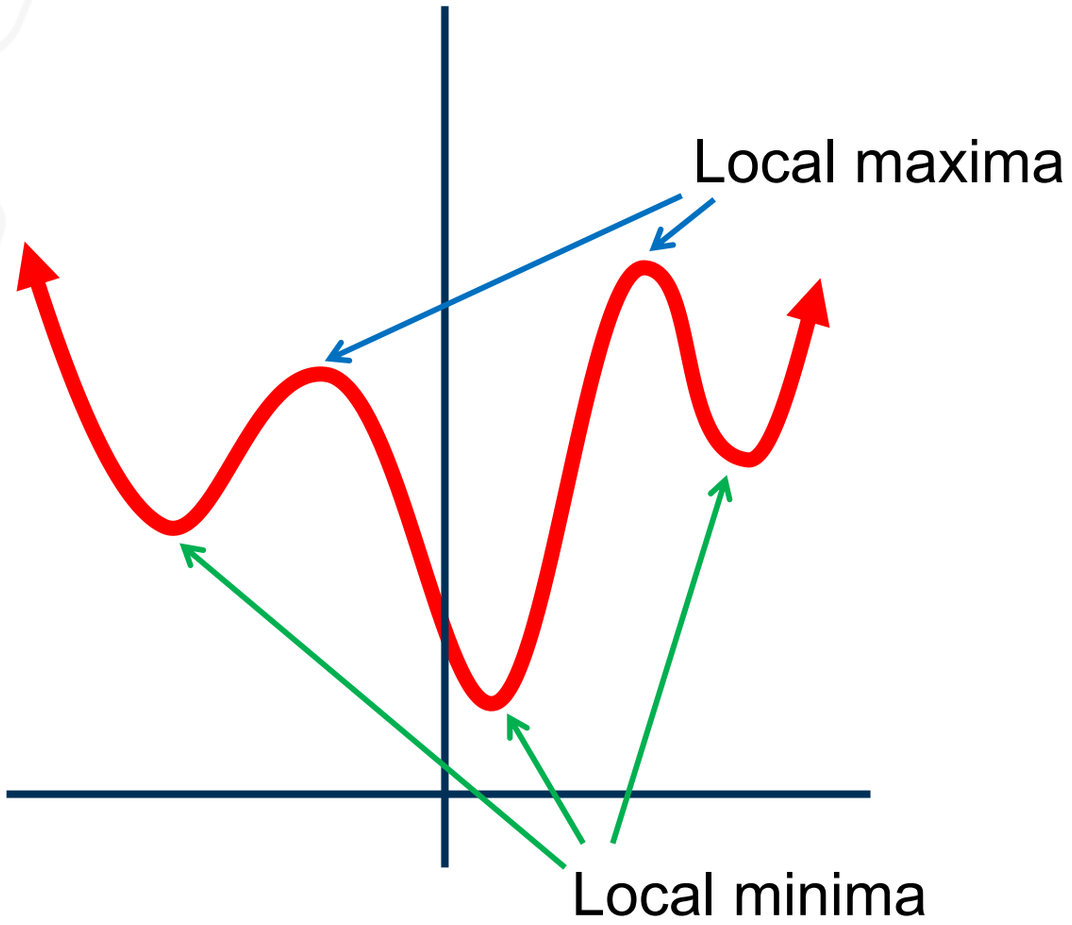
$$\leq \log 1$$

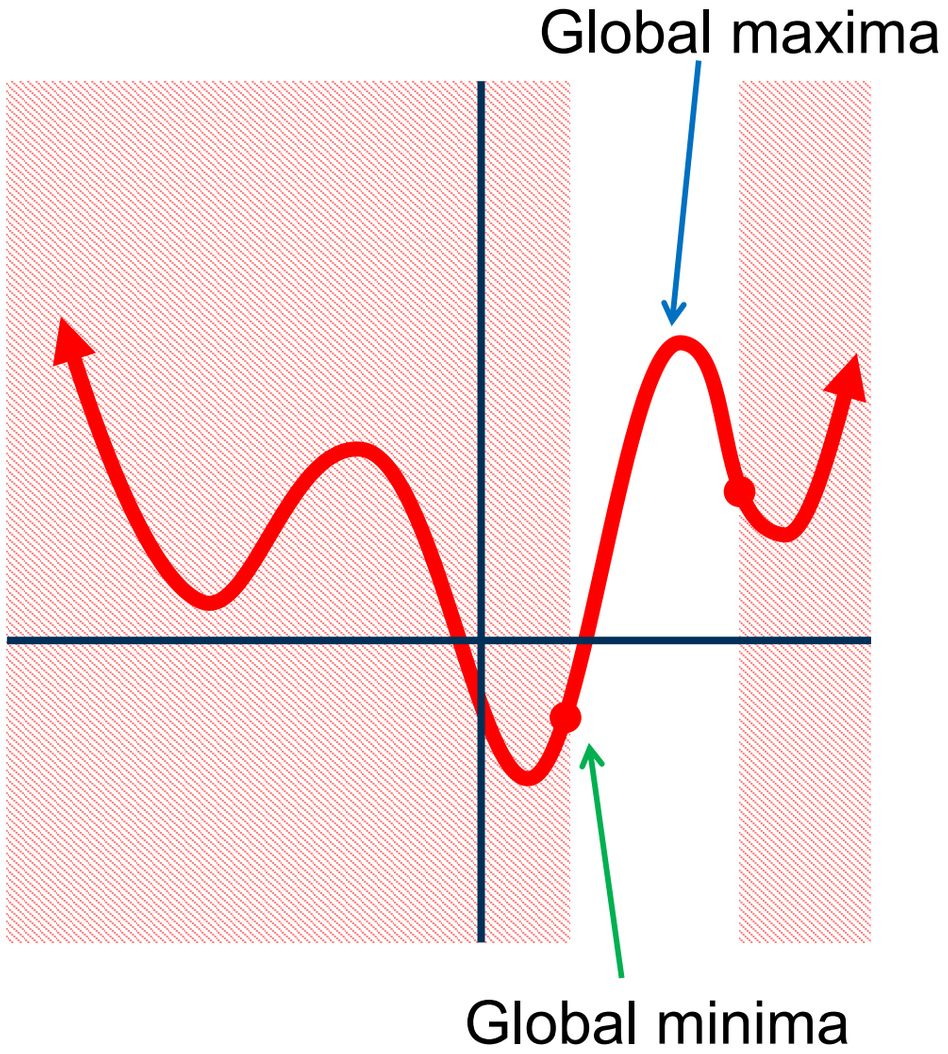
$$-KL[P][Q] \leq 0$$

$$KL[P][Q] \geq 0$$

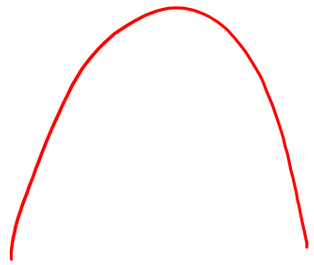
Georgia Tech

Unconstrained and constrained optimization





Basic unconstrained optimization problem



- Objective $f(\mathbf{x})$ the quantity we are trying to optimize (maximize or minimize)

- The variables $x_1, x_2, \ldots, x_m$ which can be represented in vector form as $\mathbf{x}$

(Note: the subscript m indicates a feature NOT a datapoint in our dataset)

- Finding the stationary points where the function is no longer increasing or decreasing

Equality constrained optimization

- An equality constrained optimization problem is expressed as:

$$\min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$$

$$\text{subject to } g(\mathbf{x}, \mathbf{y}) = 0$$

constraint is $x+y=5$

$g(x,y) = x+y-5$

Inequality constrained optimization

- An inequality constrained optimization problem is expressed as:

$$\min_{\mathbf{x}} f(\mathbf{x}, \mathbf{y})$$

$$\text{subject to } g(\mathbf{x}, \mathbf{y}) \leq 0$$

KKT conditions.

Linear, Quadratic, and Nonlinear Programming

Linear Programming (LP)
Objective: Linear function of x.
Constraints: Linear.
Example: $\min 3\,x_1 + 5x_2$ *s.t.* $x_1 + 2x_2 \leq 6$

Quadratic Programming (QP)
Objective: Quadratic
Constraints: Linear.
Example: $\min x^2 + y^2 + 3x + 5y$ *s.t.* $x + y \leq 4$

Nonlinear Programming (NLP)
Objective: Nonlinear.
Constraints: Can be nonlinear.
Example: $\min x^2 + y^2 + 3x + 5y$ *s.t.* $x^2 + y \leq 4$

WILL DEAL WITH THESE 2 ONLY.

Recap

- Cross Entropy

Amount of info needed to encode r.v. x with true pdf $p(x)$ but assuming to follow predicted pdf $q(x)$

$$CE = -\sum_x p(x) \log q(x) = H[P] + KL[P][Q]$$

good objective or loss function in ML ?

it leads to quicker convergence

MLE ← minimize -ve log likelihood

- Loss Functions

$f(x) \rightarrow$ objective

MSE ↓

Dot product ↑

CE ↓

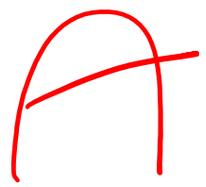
- KL Divergence

a kind of distance metric between 2 distribution

$KL[P][Q] \geq 0$

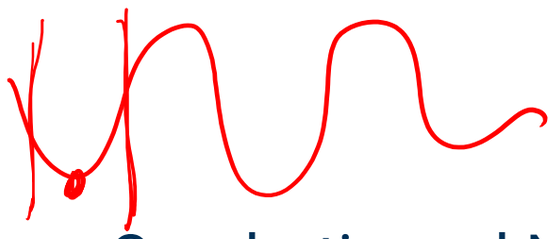
- Jensen Inequality

$E[f(x)] \geq f(E[x])$



$E[f(x)] \leq f(E[x])$

- Unconstrained and Constrained Optimization



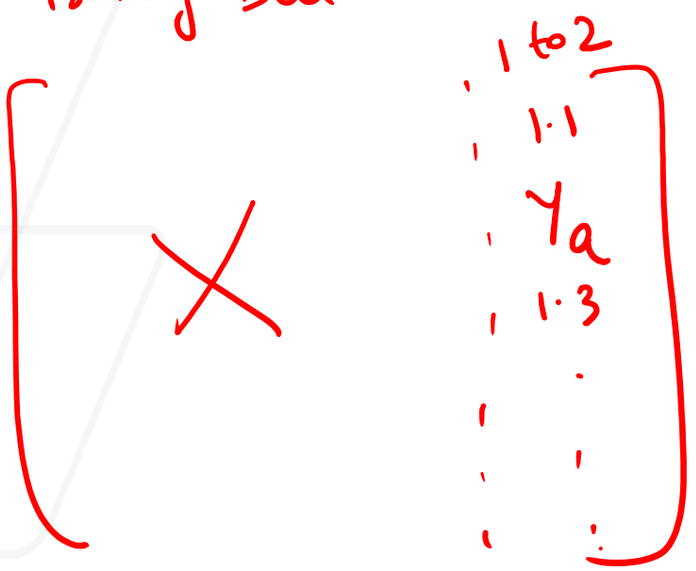
- Linear, Quadratic and Non-Linear Programming

Linear: $f(x)$ is linear, $g(x)$ is linear

Quadratic: $f(x)$ is quadratic, $g(x)$ is linear

Georgia Tech

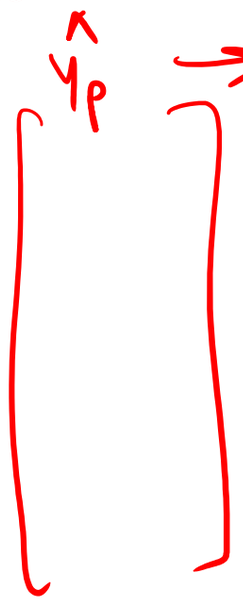
Training Dataset



Select Model

θ_{ini} = random

Regression



→ [1, 2]

Construct an objective function

$$CF = \sum_{i}^{\text{i3}} y_a \cdot \log \hat{y}_p^{\text{i3 T}}$$

CF = very high value

CF will be very small

↓

then θ is optimal

↓

model

New data → [Trained model] → $\hat{y}_p$

Equality constrained optimization problem

$\min f(x)$ s.t $g(x)=0$

$f(x,y) = x^2+y^2$

There is a notion of a Lagrangian expression where we combine both the objective function and our constraints

$g(x,y) = x+y-3000$

Lagrange expression → $L(\mathbf{x},\lambda) = f(\mathbf{x}) - \lambda g(\mathbf{x})$

objective function; lagrange multiplier; constraint

$s.t.\ \lambda \neq 0$

$g(x)=0$

$f(x)$ ← objective function needs to be optimized

$\frac{\partial L}{\partial x}=0$, $\frac{\partial L}{\partial y}=0$

$\frac{\partial f(x)}{\partial x}=0 \Rightarrow$ most optimal θ.

$f(x,y) = x^2 + y^2$ → amt. spent on rent

amt. spent on food

$s.t.\ x+y=3000$

$g(x)=0$

$\min f(x) \longrightarrow \frac{\partial f(x,y)}{\partial x}=0$, $\frac{\partial f(x,y)}{\partial y}=0$

$2x=0$

$2y=0$

$x=0, y=0$

Equality constrained optimization problem

$f(x) = x^2+y^2$ s.t. $x+y=3000$, $x+5y=1$

$L(x,y,\lambda) = x^2+y^2 - \lambda_1(x+y-3000) - \lambda_2(x+5y-1)$

$\frac{\partial L}{\partial x}=0$ $\frac{\partial L}{\partial y}=0$

- With multiple equality constraints the Lagrangian is expressed as follows:

$$L(\mathbf{x},\lambda) = f(\mathbf{x}) - \sum_{i=1}^{k} \lambda_i g_i(\mathbf{x})$$

$$s.t. \quad \lambda_i \neq 0$$

Equality constrained optimization problem

$$L(x, y, \lambda) = f(x,y) - \lambda g(x,y)$$

$$\nabla L(x, y, \lambda) = \nabla f(x,y) - \lambda \nabla g(x,y) = 0$$

$$\nabla f(x,y) = \lambda \nabla g(x,y)$$

- To solve this optimization problem, we only need to satisfy the stationarity condition, which states:

$$\nabla L(\mathbf{x}, \lambda) = 0$$

- We also need to satisfy

$$\lambda_i \neq 0$$

- This appeals to the intuition that at the optimal point, the gradient of the equality constraint is proportional to the gradient of the objective function

Equality constrained optimization: *Example 1*

$f(x,y) = x^2+y^2$

$g(x,y) = x+5y-1$

$L(x,y,\lambda) = x^2+y^2-\lambda(x+5y-1)$

$\frac{\partial L}{\partial x} = 0 \quad 2x-\lambda=0 \Rightarrow \lambda = 2x \Rightarrow x = \lambda/2$

$\frac{\partial L}{\partial y} = 0 \quad 2y-5\lambda=0 \Rightarrow \lambda = \frac{2y}{5} \Rightarrow y = 5\lambda/2$

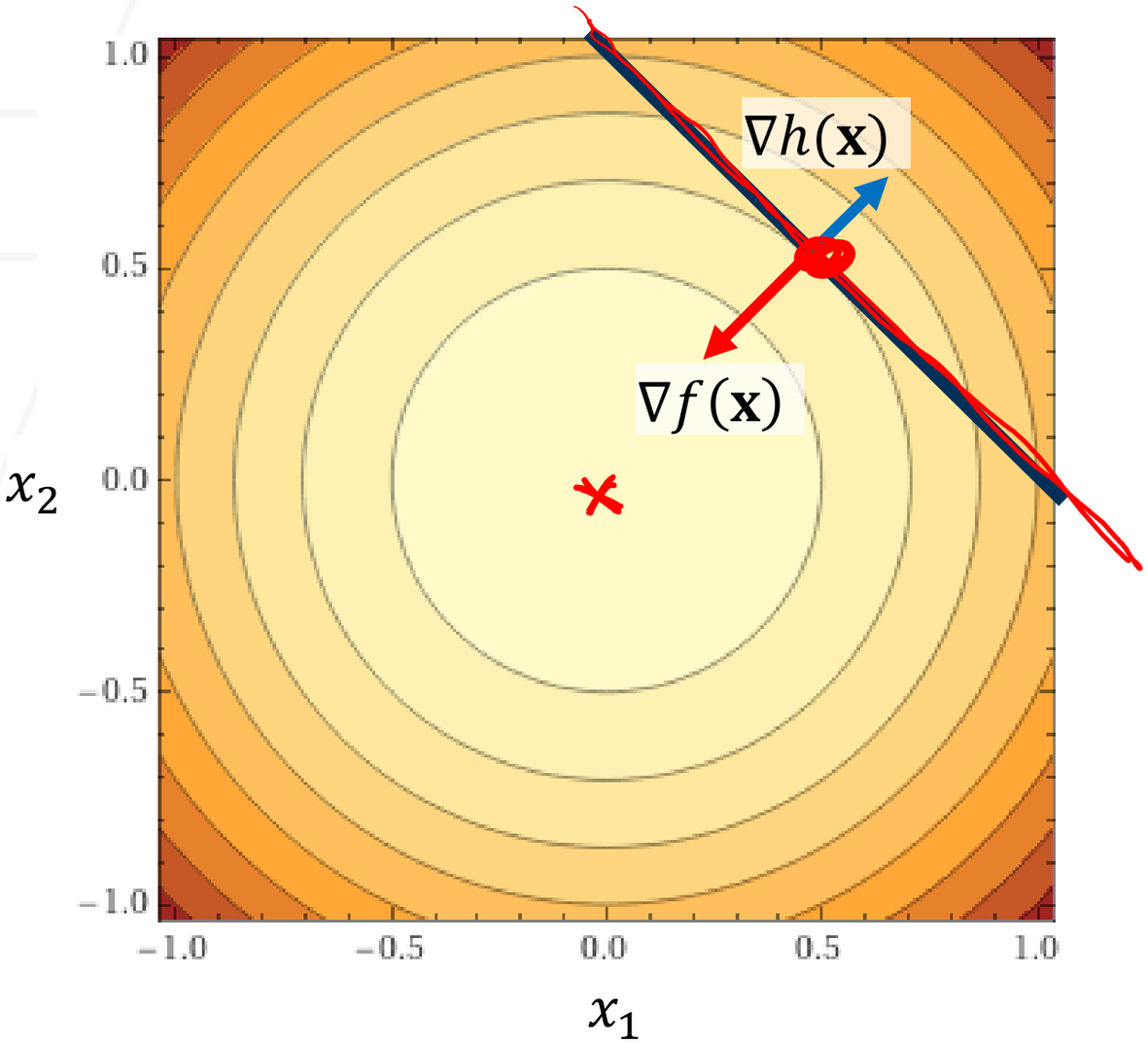
$g(x) = 0 \quad x+5y=1$

$\frac{\lambda}{2} + 5\times\frac{5\lambda}{2} = 1 \Rightarrow \frac{26\lambda}{2} = 1 \Rightarrow \lambda = 1/13$

$x = 1/26$

$y = 5/26$

Equality constrained optimization: *Example 2*



$$\max_{\mathbf{x}} \quad 1 - x_1^2 - x_2^2$$

$$s.t. \quad x_1 + x_2 = 1$$

$f(x_1, x_2) = -1 + x_1^2 + x_2^2$

$g(x_1, x_2) = x_1 + x_2 - 1$

$L(x_1, x_2, \lambda) = -1 + x_1^2 + x_2^2 - \lambda(x_1 + x_2 - 1)$

Equality constrained optimization: *Example*

$$\max_{\mathbf{x}} \quad 1 - x_1^2 - x_2^2$$

$$s.t. \quad x_1 + x_2 = 1$$

1. Write the Lagrangian for this problem

$$L(\mathbf{x}, \lambda) = 1 - x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1)$$

2. Calculate the gradient with respect to $\mathbf{x}$

$$\frac{\partial L}{\partial x_1} = -2x_1 + \lambda = 0 \rightarrow \lambda = 2x_1$$

$$\frac{\partial L}{\partial x_2} = -2x_2 + \lambda = 0 \rightarrow \lambda = 2x_2$$

Equality constrained optimization: *Example*

3. Solve for $\mathbf{x}^*$ and λ

Using our previous two derivatives and the equality constraint, we obtain:

$$x_1 + x_2 = 1 \rightarrow x_1 + x_1 = 1 \rightarrow x_1^* = x_2^* = \frac{1}{2}$$

From $\frac{\partial L}{\partial x_1} = 0$ we find the value of λ

$$\lambda = 2x_1 \rightarrow \lambda = 1$$

Inequality constrained optimization problem

- Constraints that limit how small or big variables can be. These are inequality constraints noted as $g_j(\mathbf{x})$

- An optimization problem is usually expressed as:

$$\min_{\mathbf{x}} f(\mathbf{x})$$

objective

$$\text{subject to } g_j(\mathbf{x}) \leq 0, j = 1, 2, \ldots, m$$

.

constraint

Inequality constrained optimization problem

↳ we have to add $\lambda g(x)$

- The optimization problem can be rewritten as

$$\min_{\mathbf{x}} \; f(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j g_j(\mathbf{x}) \qquad \text{subject to} \quad \boldsymbol{\lambda} \geq 0$$

≥ 0 ≤ 0

- We need to satisfy the Karush-Kuhn-Tucker (KKT) conditions on $\mathbf{x}$, $\boldsymbol{\lambda}$ which is **necessary and sufficient** for optimality

Karush-Kuhn-Tucker Conditions

- Stationarity condition:

$$\min L(\mathbf{x}, \lambda) = f(\mathbf{x}) + \sum_{j=1}^{m} \lambda_j g_j(\mathbf{x})$$

- Primal feasibility

$$g_j(\mathbf{x}) \leq 0, j = 1, 2, \ldots, m$$

- Dual Feasibility

$$\lambda_j \geq 0$$

- Complementary slackness

$$\lambda_j g_j(\mathbf{x}) = 0, j = 1, \ldots, m$$

If g(x) < 0 → **inactive** solution since λ =0
If g(x) = 0 → **active** solution since λ is at play and > 0, which means constraint is active

Karush-Kuhn-Tucker Interpretations

- The complementary slackness condition applies only to inequality constraints
- It states that for a given inequality constraint…

$$\ldots\text{either } g_j(x) = 0 \quad \text{or} \quad \lambda_j = 0$$

- Whenever the dual variable $\lambda_j > 0$ and therefore $g_i(x) = 0$, we say that the **constraint is active** or that the **constraint is tight at x**
- The reverse $\lambda_j = 0$ and therefore $g_j(x) < 0$ means that the constraint is inactive, meaning that it is not effectively impacting the solution

Rule: For each inequality constraint, **either** it is active and has a positive multiplier, or it is inactive and its multiplier is zero.

Primal vs. Dual Form

Quadratic Prog. is popular because it leads to strong duality

Primal Form

$$f(x,y) = \frac{x^2}{2} + \frac{y^2}{2} \qquad g(x,y) = x + y - 24$$

$$L(x,y,\lambda) = \frac{x^2}{2} + \frac{y^2}{2} - \lambda(x+y-24)$$

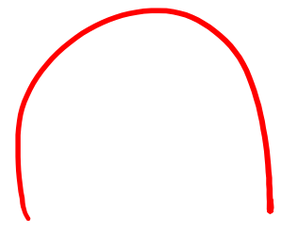
$$\frac{\partial L}{\partial x} = 0 \qquad x - \lambda = 0 \qquad x = \lambda$$

$$\frac{\partial L}{\partial y} = 0 \qquad y - \lambda = 0 \qquad y = \lambda$$

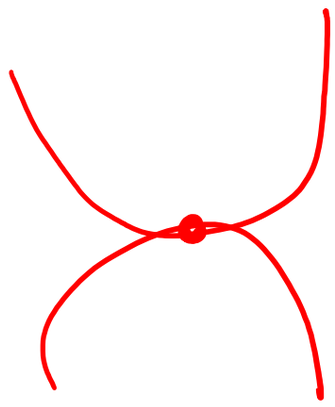
Dual Form

$$L(\lambda) = \frac{\lambda^2}{2} + \frac{\lambda^2}{2} - \lambda(\lambda+\lambda-24)$$

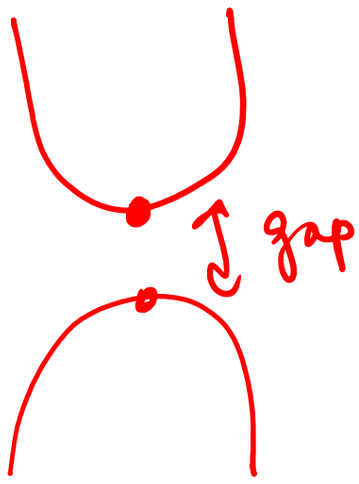
$$= \lambda^2 - \lambda^2 - \lambda^2 + 24\lambda = -\lambda^2 + 24\lambda$$



$$\frac{\partial L}{\partial \lambda} = 0$$



Strong Duality



Weak duality

Primal vs Dual Form

Primal Problem
Original optimization problem

Dual Problem
Derived problem as **a function of** λ only.

Analogy:
"Instead of directly solving for x and y together, we solve for what x and y *should look like* given a certain λ (like a candidate penalty price). Then we search over λ to find the best price that satisfies the constraint. This flips a 2D search into a 1D search."

Motivation:
In big problems (many x's), solving for x in terms of λ can be done analytically or numerically, but then the dual problem might be much lower-dimensional → computationally cheaper.

Primal vs Dual Form

Weak Duality
- Dual optimum ≤ primal optimum (for minimization problems).
- Always true.

Strong Duality
- Dual optimum = primal optimum.
- Holds for **convex problems** with feasible solution.

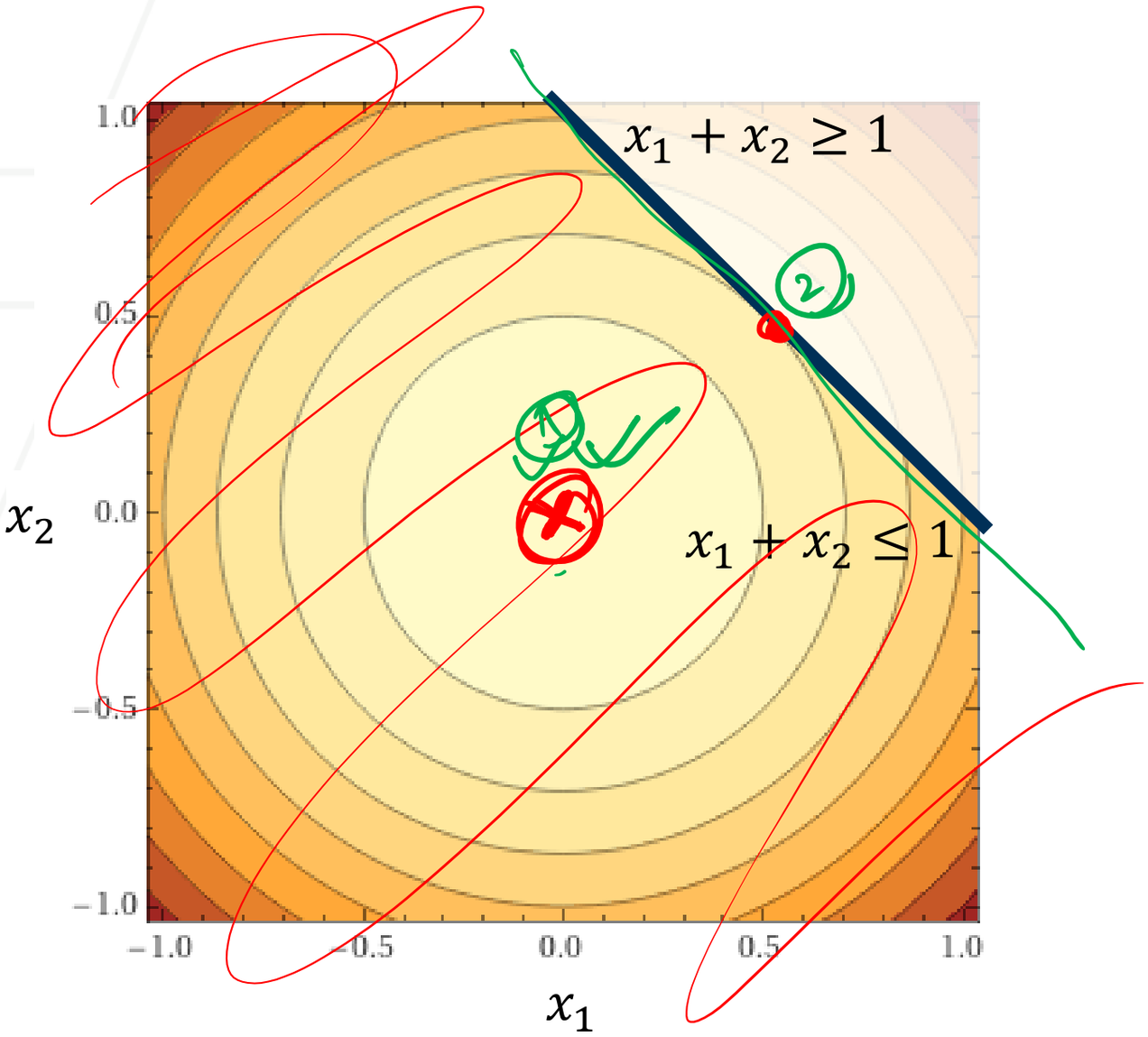
This is why **Quadratic Programming** is popular: QP is convex and has efficient solvers that work on the dual.

Why care?
Many ML models (e.g., SVMs) are solved via the dual.

Inequality constrained optimization: *Example 1*

$f(x) = x^2+y^2$ s.t. $x+y \geq 1$

$g(x,y) = -x-y+1 \leq 0$



$$\max_{\mathbf{x}} \quad 1 - x_1^2 - x_2^2$$

$$s.t. \quad x_1 + x_2 \leq 1$$

$g(x_1,x_2) = x_1+x_2-1 \leq 0$

Georgia Tech

$g(x) \leq 0$
$\lambda \geq 0$

Inequality constrained optimization: *Example 1*

1. Write the Lagrangian for this problem

$$L(\mathbf{x}, \lambda) = -1 + x_1^2 + x_2^2 + \lambda(x_1 + x_2 - 1)$$

$f(x)$

2. Calculate the gradient with respect to $\mathbf{x}$ (stationarity condition)

$$\frac{\partial L}{\partial x_1} = 2x_1 + \lambda = 0 \rightarrow \lambda = -2x_1$$

$$\frac{\partial L}{\partial x_2} = 2x_2 + \lambda = 0 \rightarrow \lambda = -2x_2$$

$$x_1 = x_2$$

Inequality constrained optimization: *Example 1*

3. Use the complementary slackness

Case 1: $\lambda > 0$ and $g(x) = 0$

Active Sol.

$$x_1 + x_2 - 1 = 0 \rightarrow 2x_1 = 1$$

$$x_1 = x_2 = \frac{1}{2}$$

4a. Check if the dual feasibility condition is satisfied

$$\lambda = -2x_1 = -1$$

Dual feasibility X

This solution **does NOT satisfy the dual feasibility condition** and it is thus **unfeasible**

Inequality constrained optimization: *Example 1*

3. Use the complementary slackness

Case 2: $\lambda = 0$ and $g(x) < 0$

$$\lambda = -2x_1$$

$$x_1 = x_2 = 0$$

4b. Check if the primal feasibility condition is satisfied:

$$g(x_1, x_2) = 0 + 0 - 1 = -1 < 0$$

The solution is **feasible** and therefore the optimal!

Inequality constrained optimization: *Example 1*

Summary of the KKT conditions for the problem:

$$\left.\begin{aligned}\lambda &= -2x_1\\ \lambda &= -2x_2\end{aligned}\right\}\ \textbf{Stationarity}$$

$$x_1 + x_2 - 1 \leq 0 \quad \textbf{Primal feasibility}$$

$$\lambda \geq 0 \quad \textbf{Dual feasibility}$$

$$\lambda(x_1 + x_2 - 1) = 0 \quad \textbf{Complementary slackness}$$

Gradient descent

$f(x) = e^x + 2x$

← Numerical Approaches

Gradient Ascent

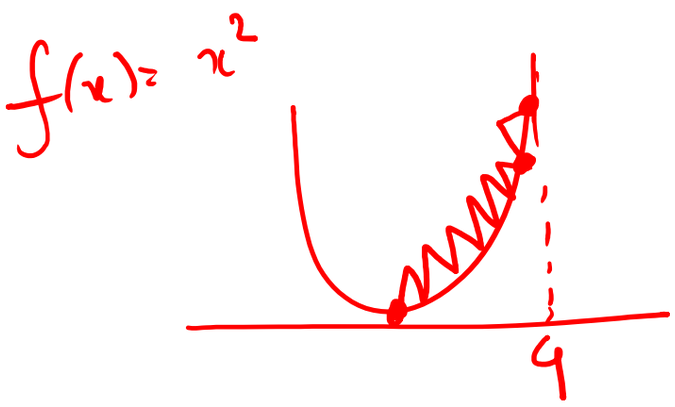
- In machine learning, some of the optimization problems suffer from the following challenges:

$x_{n+1} = x_n + \alpha \nabla f(x_n)$

  - A closed form solution may be computationally intractable
    - Functions are complex: no easy derivative = 0 solution.
    - High dimensions: solving $\nabla f(x)=0$ analytically is impossible.
  - All the data may not be available at the time of training (e.g. streaming data)

- Gradient descent is an iterative minimization technique for differentiable functions on a domain
- **Intuition:** the function $f(\mathbf{x})$ decreases the fastest by going from x_n to x_{n+1} in the opposite direction of the gradient of $f(\mathbf{x})$

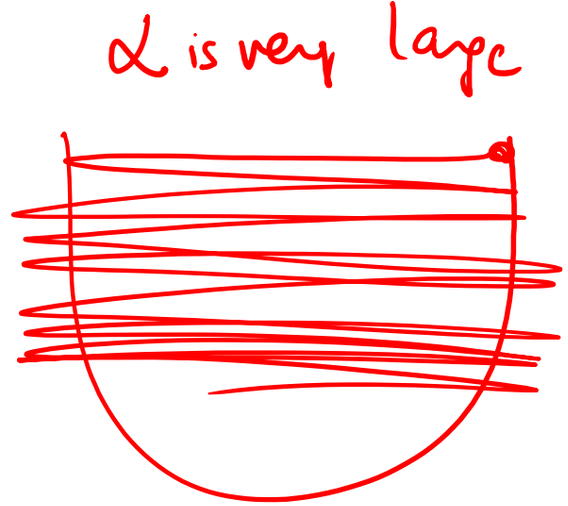
$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha_n \nabla f(\mathbf{x}_n)$$



$x_0 = 4 \quad x_1 = x_0 - \alpha_0 \nabla f(x_0)$

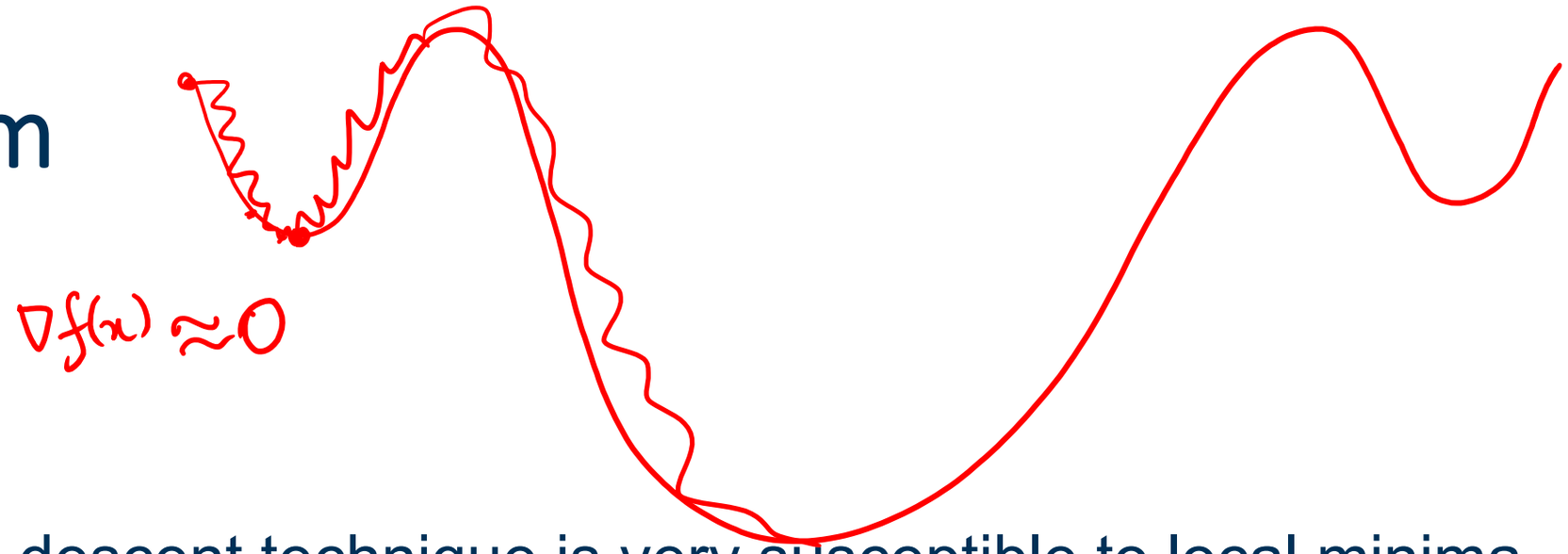
learning rate
HYPERPARAMETER

Learning rate



- The choice of the learning rate α_n plays an important role in the success of the minimization:
  - If α_n is **too small** then the number of iterations necessary to reach the minimum will be too large
  - If α_n is **too large** then we may overshoot the minimum
- It can be set to a constant or varied iteratively to improve the performance of the algorithm

Momentum



- The gradient descent technique is very susceptible to local minima, therefore, it we can modified by adding a momentum term

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \alpha_n g(\mathbf{x}_n)$$

Where:

$$g(\mathbf{x}_n) = \mu g(\mathbf{x}_{n-1}) + \nabla f(\mathbf{x}_n)$$

μ = momentum constant

- The update direction is a running average of past gradients.
- The momentum constant also needs to be fine tuned along with the learning rate to each problem, typically picked as a value between 0 and 1

Quick Knowledge Check

$E[f(x)] \geq f(E[x])$

- Which property of a function is required to apply Jensen's inequality? A. Differentiability B. Convexity or concavity C. Existence of a global minimum D. Finite domain

- When is an inequality constraint considered **inactive** at the solution? A. When the constraint holds with equality B. When $\lambda>0$ C. When the unconstrained optimum satisfies the constraint D. When the gradient is zero

- Which KKT condition ensures that either the constraint is tight or the multiplier is zero? A. Primal feasibility B. Dual feasibility C. Stationarity D. Complementary slackness

$g(w) \leq 0$ $\lambda \geq 0$ $\nabla L = 0$

- What type of optimization landscape most strongly motivates the use of **momentum**? A. Linear objectives B. Perfectly symmetric bowls C. Narrow valleys and shallow local minima

Georgia Tech