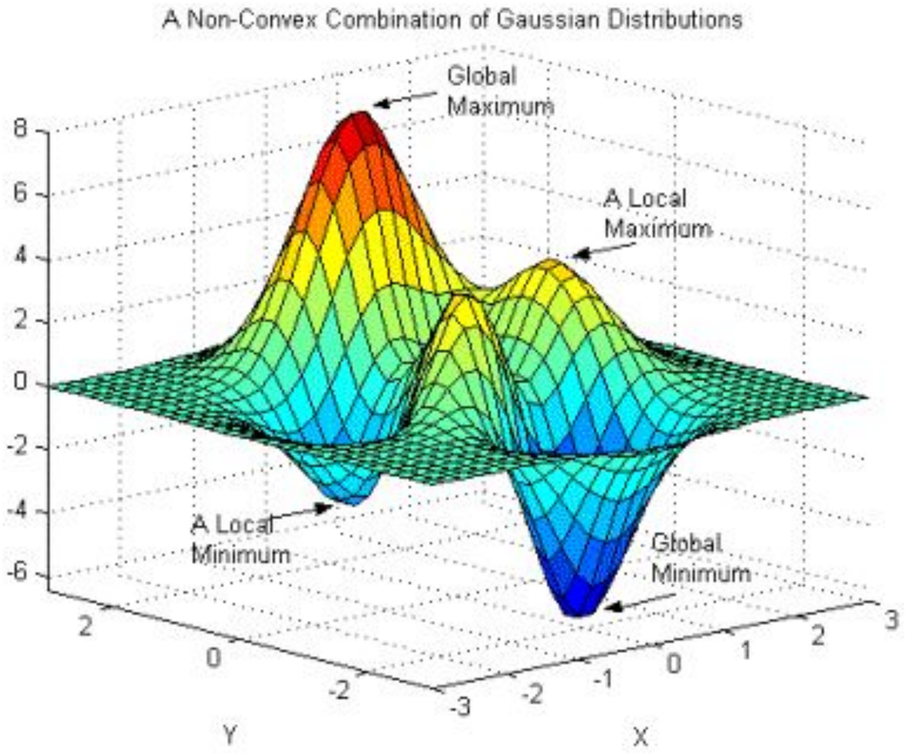
10-315 Recitation #2

Convexity & Optimization

What is optimization?

- Different kinds of optimization problems in mathematics
  - LPs, IPs, zeroes and optima of functions
- In this class we're mostly concerned with finding local and global optima
  - Coordinate descent, **gradient descent**, interpolating polynomials (later on in class)



Gradients

- Definition: $\nabla f(\mathbf{x}) = \langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_n} \rangle f(x)$

Partial derivative: taking the derivative with respect to one variable

- Simplifying assumption: variables are not dependent on each other, so derivative of x_2 with respect to x_1 is 0

Example: let $f(\mathbf{x}) = \frac{x_1^2}{2} + \frac{x_2^2}{2}$

$\nabla f(\mathbf{x}) =$ [redacted]

What is the gradient of f at $(1, 2)$?

$\nabla f(1, 2) =$ [redacted]

Gradients

- Definition: $\nabla f(\mathbf{x}) = \langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, ..., \frac{\partial}{\partial x_n} \rangle f(x)$

Partial derivative: taking the derivative with respect to one variable

- Simplifying assumption: variables are not dependent on each other, so derivative of x_2 with respect to x_1 is 0

Example: let $f(\mathbf{x}) = \frac{x_1^2}{2} + \frac{x_2^2}{2}$

$\nabla f(\mathbf{x}) = \langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2} \rangle \cdot (\frac{x_1^2}{2} + \frac{x_2^2}{2}) = \langle \frac{\partial}{\partial x_1}(\frac{x_1^2}{2} + \frac{x_2^2}{2}), \frac{\partial}{\partial x_2}(\frac{x_1^2}{2} + \frac{x_2^2}{2}) \rangle = \langle (x_1 + 0), (0 + x_2) \rangle = \langle x_1, x_2 \rangle$

What is the gradient of f at $(1, 2)$?

$\nabla f(1, 2) =$ [redacted]

Gradients

- Definition: $\nabla f(\mathbf{x}) = \langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_n} \rangle f(x)$

Partial derivative: taking the derivative with respect to one variable

- Simplifying assumption: variables are not dependent on each other, so derivative of x_2 with respect to x_1 is 0

Example: let $f(\mathbf{x}) = \frac{x_1^2}{2} + \frac{x_2^2}{2}$

$\nabla f(\mathbf{x}) = \langle \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2} \rangle \cdot (\frac{x_1^2}{2} + \frac{x_2^2}{2}) = \langle \frac{\partial}{\partial x_1}(\frac{x_1^2}{2} + \frac{x_2^2}{2}), \frac{\partial}{\partial x_2}(\frac{x_1^2}{2} + \frac{x_2^2}{2}) \rangle = \langle (x_1 + 0), (0 + x_2) \rangle = \langle x_1, x_2 \rangle$

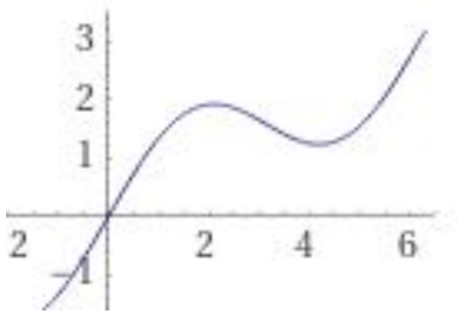
What is the gradient of f at $(1, 2)$?

$\nabla f(1, 2) = \langle x_1, x_2 \rangle (1, 2) = \langle 1, 2 \rangle$

- The gradient is a vector giving the rate of change in function value with respect to each variable
- An intuitive way to think about the gradient is as the vector that gives the direction of fastest increase
- https://www.geogebra.org/3d?lang=en -- (f, (1, 2, 2.5))
- So we see the gradient shows us the direction of fastest increase, but what if we wanted to go backwards, towards the minimum?

Gradient Descent Algorithm

- Travel in reverse direction -- the direction of greatest decrease
- Update rule: x_new = x_old – η * ∇ f(x_old)
- How far should we travel in each step given that we don’t know where the minimum is?
    - Learning rate denoted by eta (η)
- https://suniljangirblog.wordpress.com/2018/12/03/the-outline-of-gradient-descent/ (visualized)
- Choice of learning rate can be very important
- Definition of convergence for solvers
- Algorithm relies on convexity

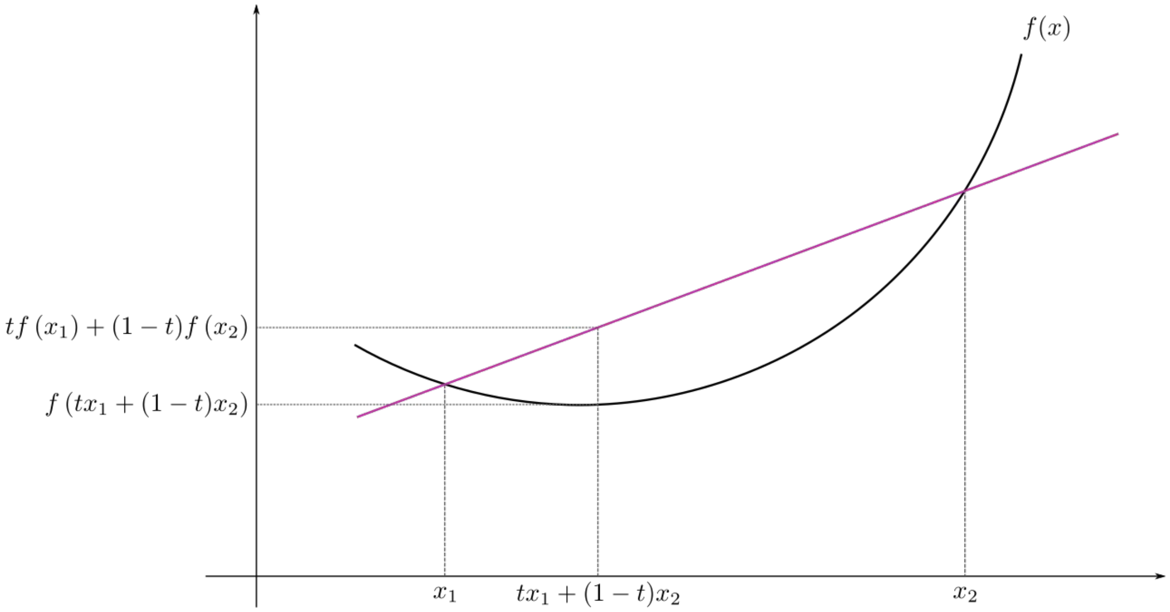


Convexity

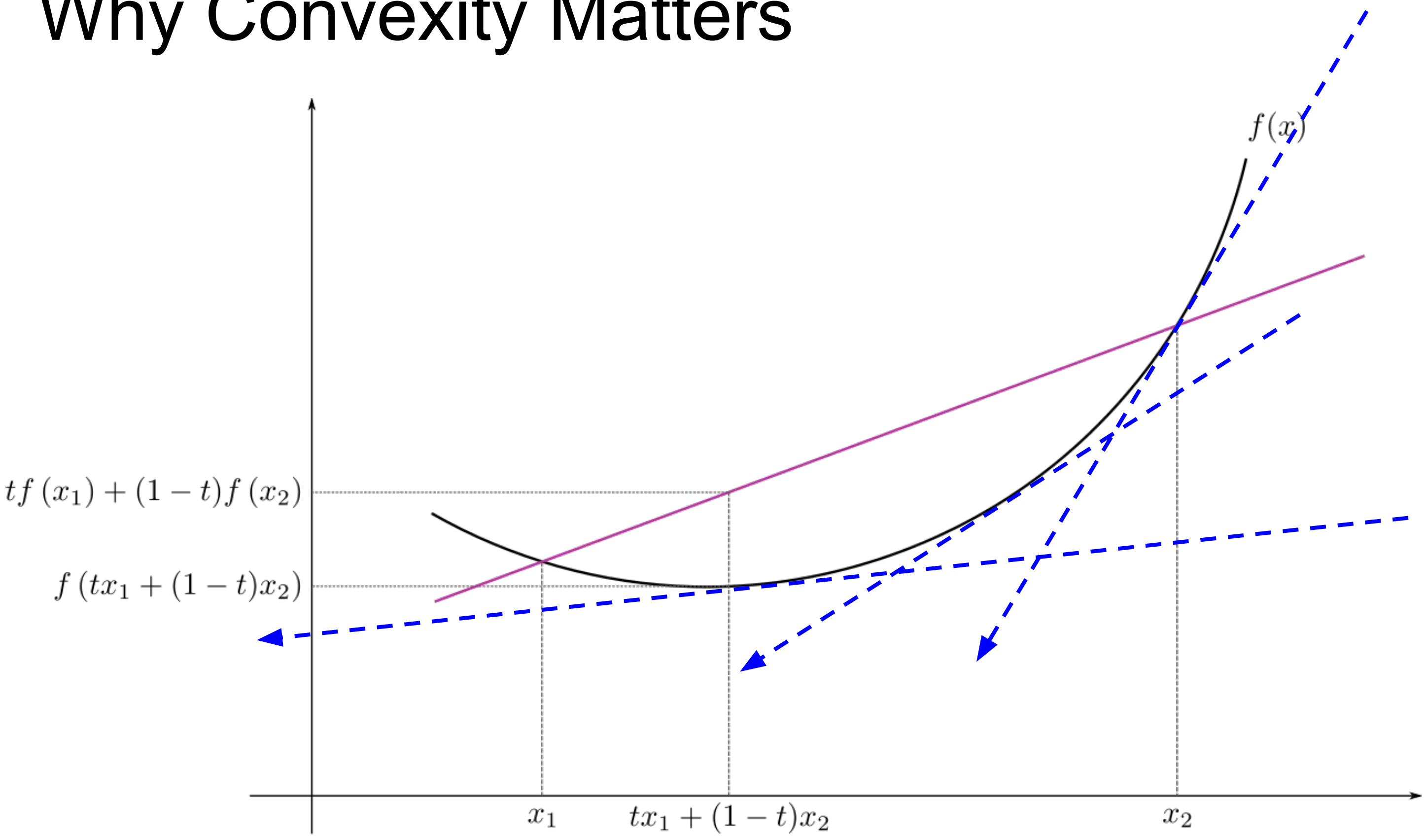
f is called **convex** if:

$$\forall x_1, x_2 \in X, \forall t \in [0,1]: \qquad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

- But why does convexity matter for optimization?



Why Convexity Matters



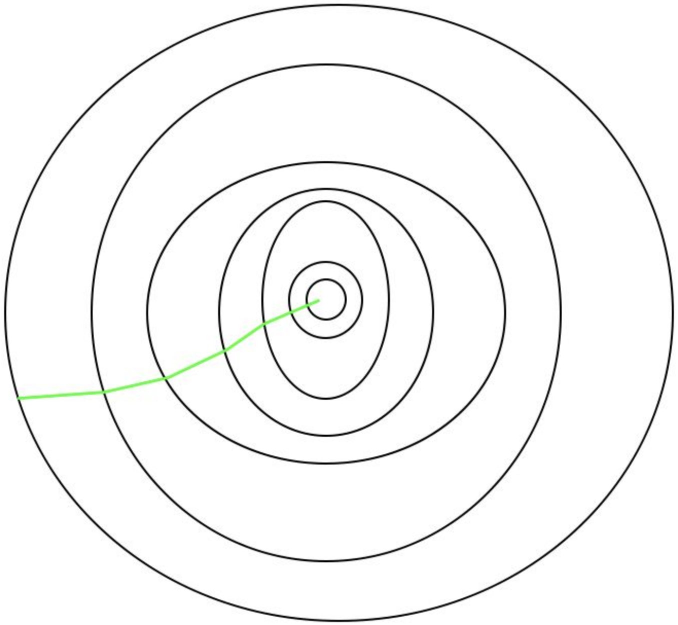
- Convexity guarantees that gradient descent will approach an optimum
- Without this guarantee, gradient descent may never converge

Stochastic Gradient Descent

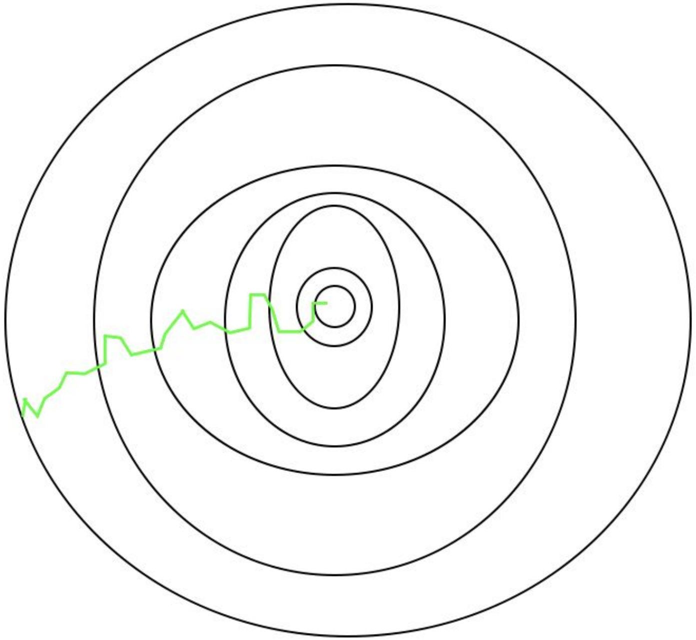
- Normal gradient descent uses batches of data (often the entire dataset) to determine the gradient in each step
- For large datasets this can be very expensive
- We can also randomly select one data point at each iteration to use for computing the gradient
- This will be less accurate at each step, but in expectation each step should still be towards the optimum

Normal GD vs. SGD

batch-based GD



single sample SGD



Example Problems

Compute the gradient of this function: $f(x , y) = x^2 + 2y^2$

1. Starting at the point (4, 1), run four iterations of gradient descent using the learning parameter $\eta = 0.25$.
2. Starting at the point (6, 2), run four iterations of gradient descent using the learning parameter $\eta = 0.5$.
3. Let $f(x, y) = 1.783(x-2)^2 + 2.481(y+3)^2$. Starting at the point (37.4, 90.2), run gradient descent using the learning parameter $\eta = 0.1$ until you get within 0.001 of the function minimum.

*update rule: $x\_new = x\_old – \eta * \nabla f(x\_old)$

Conditional Independence

A and B are conditionally independent given C if $P(A \cap B|C) = P(A|C)P(B|C)$
Equivalently, A and B are conditionally independent given C if $P(A|B \cap C) = P(A|C)$

- Knowing that C has occurred, A and B have no impact on each other
- Not the same as regular independence
- Regular independence implies conditional independence, converse is not true
- Important in ML -- we assume data rows are conditionally independent given some set of parameters
  - Each row is some observation from a distribution. We assume these observations are independent given the underlying parameters (example in next slide)

Concavity of Bernoulli Likelihood

$L(\theta) = p(X_1, X_2, ..., X_n|\theta)$

[redacted]

Concavity of Bernoulli Likelihood

$$L(\theta) = p(X_1, X_2, ..., X_n|\theta)$$
$$= p(X_1|\theta)p(X_2|\theta)...p(X_n|\theta)$$
$$= \Pi_{i=1}^{n} p(X_i|\theta)$$
$$= \Pi_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$$

[redacted]

Concavity of Bernoulli Likelihood

$$L(\theta) = p(X_1, X_2, ..., X_n|\theta)$$
$$= p(X_1|\theta)p(X_2|\theta)...p(X_n|\theta)$$
$$= \Pi_{i=1}^{n} p(X_i|\theta)$$
$$= \Pi_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$$
$$\Rightarrow \log(L(\theta)) = \sum_{i=1}^{n} \log \theta^{X_i}(1-\theta)^{1-X_i}$$
$$= \sum_{i=1}^{n} X_i \log \theta + (1 - X_i) \log(1-\theta)$$
$$= (\sum_{i=1}^{n} X_i) \log \theta + (n - \sum_{i=1}^{n} X_i) \log(1-\theta)$$

[redacted]

Concavity of Bernoulli Likelihood

$L(\theta) = p(X_1, X_2, ..., X_n|\theta)$
$= p(X_1|\theta)p(X_2|\theta)...p(X_n|\theta)$
$= \Pi_{i=1}^{n} p(X_i|\theta)$
$= \Pi_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i}$
$\Rightarrow \log(L(\theta)) = \sum_{i=1}^{n} \log \theta^{X_i}(1-\theta)^{1-X_i}$
$= \sum_{i=1}^{n} X_i \log\theta + (1-X_i)\log(1-\theta)$
$= (\sum_{i=1}^{n} X_i)\log\theta + (n - \sum_{i=1}^{n} X_i)\log(1-\theta)$
$\Rightarrow \frac{\partial}{\partial\theta}\log(L(\theta)) = \frac{1}{\theta}\sum_{i=1}^{n} X_i - (n - \sum_{i=1}^{n} X_i)\frac{1}{1-\theta}$
$\Rightarrow \frac{\partial^2}{\partial\theta^2}\log(L(\theta)) = -\frac{1}{\theta^2}\sum_{i=1}^{n} X_i - (n - \sum_{i=1}^{n} X_i)\frac{1}{(1-\theta)^2}$
But we know $\theta \in (0,1)$, $0 \leq \sum_{i=1}^{n} X_i \leq n$
So we conclude $\frac{\partial^2}{\partial\theta^2}\log(L(\theta)) < 0$
$\Rightarrow$ the Bernoulli likelihood function is concave down.