The general paradigm of supervised learning

- The goal of a large family of machine learning is to minimize the prediction errors of the model
  - Ideally we want to predict the true errors, errors made by the model when it is used in realistic scenarios
  - That is hard to do, so the common practice is to minimize the errors in a training sample
  - To do that we need to define an *loss function*, which is a metric that measures the errors in the prediction of the model.
    - Cross-entropy Loss, Squared Error Loss, Hinge Loss
- In other cases it is more natural to think of the goal of learning is to optimize an *objective function*, e.g., Maximum Likelihood
- Whether to call is a loss function or objective function, there is no difference in how they are optimized

Commonly used loss and objective functions in NLP

- Naïve Bayes: maximize the joint probability of a training set of labeled samples

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}}\, p(\boldsymbol{x}^{1:N}, y^{1:N}; \boldsymbol{\theta})$$

- Logistic Regression: The weights are estimated by **Maximum Conditional Likelihood**

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \log p(y^{1:N} | \boldsymbol{x}^{1:N}; \boldsymbol{\theta})$$

- SVM: The weights are estimated by minimizing marginal loss

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \sum_{i=1}^{N} \left(1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})\right)_+$$

Note: Letters in bold indicates vector: $\boldsymbol{\theta}$, $\boldsymbol{x}$, $\boldsymbol{f}$. Alternative notations: $\vec{\theta}$, $\vec{x}$, $\vec{f}$

Naïve Bayes Objective

- Naïve Bayes: Maximize the joint probability of a training set of labeled samples, in a process called **Maximum Likelihood Estimation**

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} P(\boldsymbol{x}^{1:N}, y^{1:N}; \boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \prod_{i=1}^{N} p(\boldsymbol{x}^i, y^i; \boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \log P(\boldsymbol{x}^i, y^i; \boldsymbol{\theta})$$

Logistic Regression Objective

- Logistic Regression: The weights are estimated by **Maximum Conditional Likelihood**

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} \log p(y^{1:N} | \boldsymbol{x}^{1:N}, \boldsymbol{\theta})$$

$$= \operatorname*{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y \in \mathcal{Y}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)\right) \right)$$

or by minimizing the **logistic loss**:

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} - \sum_{i=1}^{N} \left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}) - \log \sum_{y \in \mathcal{Y}} \exp\left(\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y)\right) \right)$$

Support Vector Machine Objective

- SVM: The weights are estimated by minimizing marginal loss

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} \left(1 - \gamma(\boldsymbol{\theta}; \boldsymbol{x}^{(i)}, y^{(i)})\right)_+$$

$$= \operatorname*{argmin}_{\boldsymbol{\theta}} \sum_{i=1}^{N} (\max_{y \in \mathcal{Y}} (\boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y) + c(y^{(i)}, y)) - \boldsymbol{\theta} \cdot \boldsymbol{f}(\boldsymbol{x}^{(i)}, y^{(i)}))_+$$

These look rather daunting, don't they?

How do we minimize a function?

In order to minimize a function, we need to be able to compute the *derivative*, or rate of change of the function.

Let's start with a much simpler function $f(x) = x^2 + 1$, and its derivative is:

$$\frac{d}{dx} f(x) = \frac{d}{dx}(x^2 + 1) = 2x$$

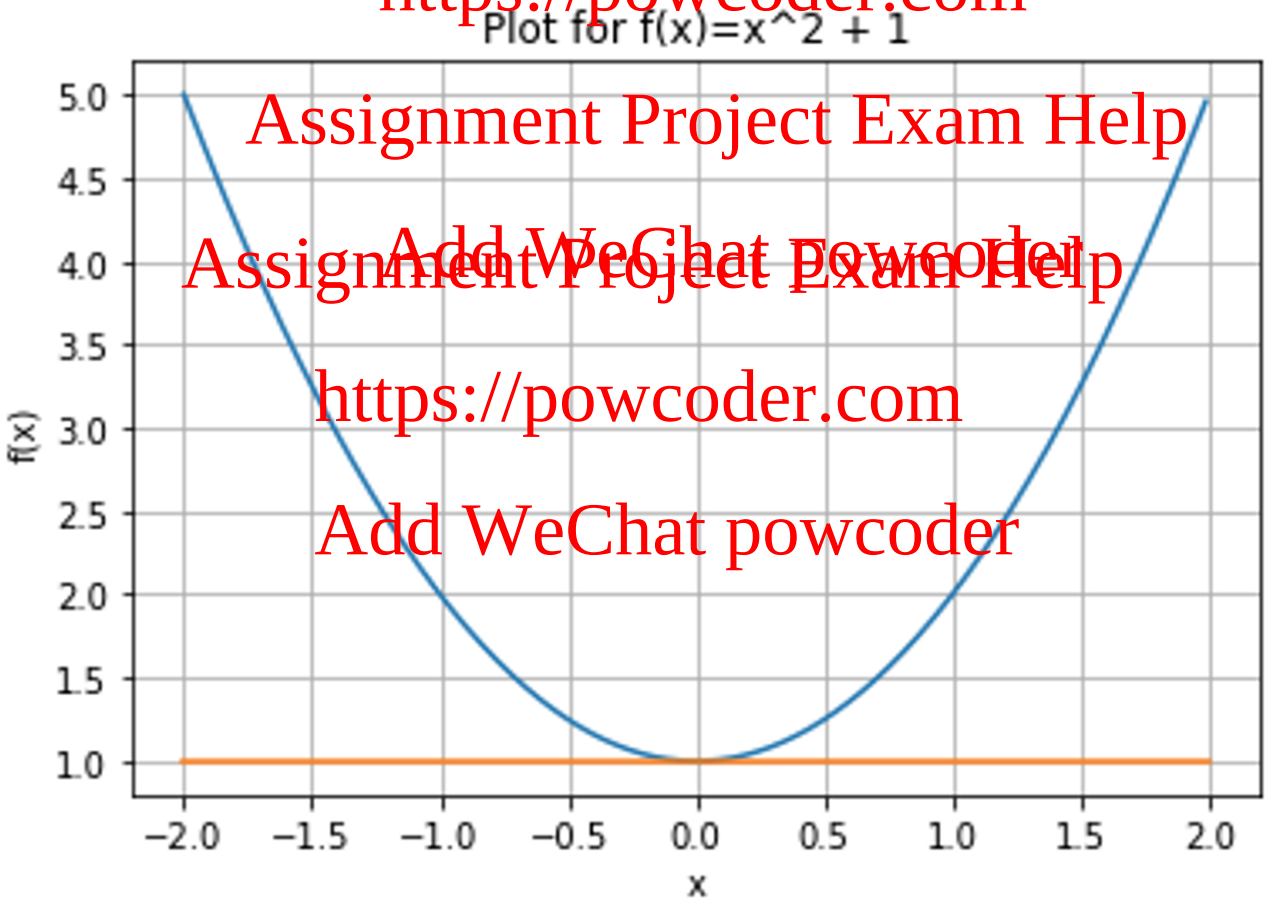
"The derivative of the function $f(x)$ with respect to (w.r.t.) x"

This looks like magic, but it's really just calculus.

How do we find the minimum of a function with the derivative?

- The derivative of a function can be interpreted as a slope at a certain point of the function.
- At the point that is the minimum (or maximum) of the function, the slope is level.
  We can find the minimum of the function by setting its derivative to zero:
  $2x = 0, x = 0$
- For this particular function, there is a closed form solution. Most models in NLP don't have a closed form solution, but some do, e.g., Naïve Bayes.

Plot the function



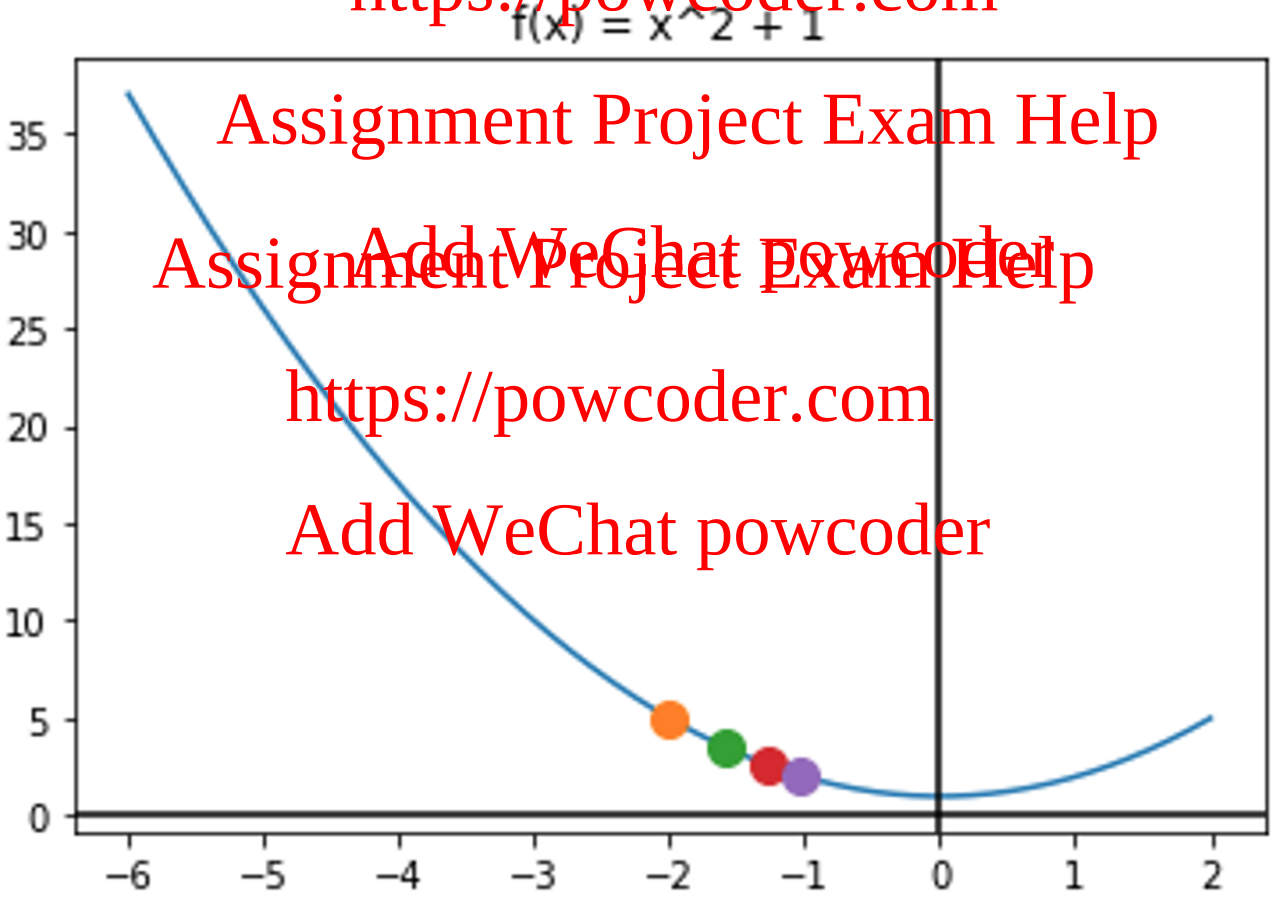
Finding the minimum iteratively

For functions that don't have a closed form solution, we find its minimum iteratively. We subtract (a fraction of) the derivative from the input x so that the value of the function will decrease. Suppose we start at the point where $x = -1$, and set the fraction $\eta \triangleq 0.1$, and $\Delta x = \eta \frac{d}{dx} f(x)$. So:

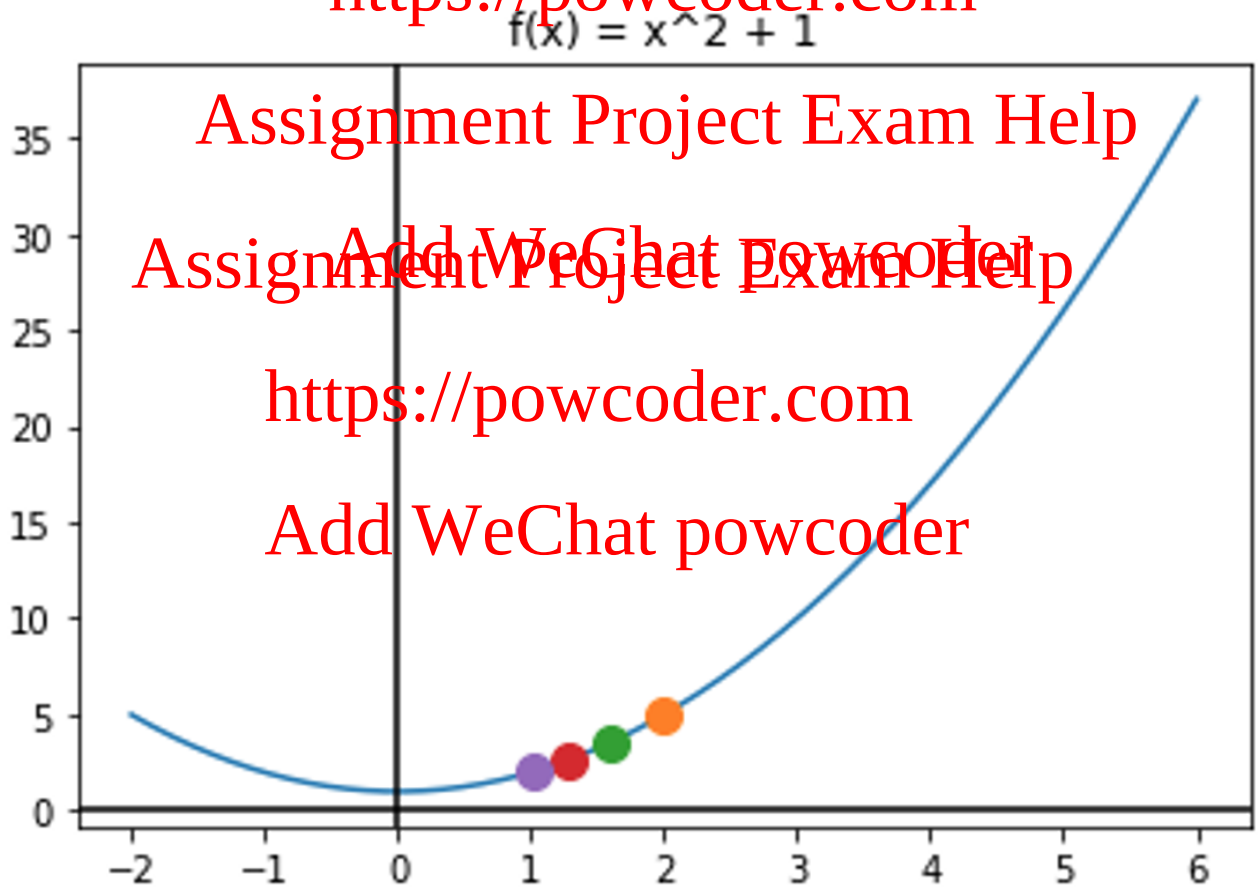
$$x = x - \Delta x = -1 - 0.1 \times (-2) = -0.8$$
$$f(x) = (-0.8)^2 + 1 = 1.64$$
$$x = x - \Delta x = -0.8 - 0.1 \times (-1.6) = -0.64$$
$$f(x) = (-0.64)^2 + 1 = 1.4096$$
$$x = x - \Delta x = -0.64 - 0.1 \times (-1.28) = -0.512$$
$$f(x) = (-0.512)^2 + 1 = 1.262144$$

As x approaches 0, $f(x)$ reaches the minimum, which is 1.

Finding the minimum iteratively



Finding the minimum iteratively



What if we try to learn fast using a larger learning rate?

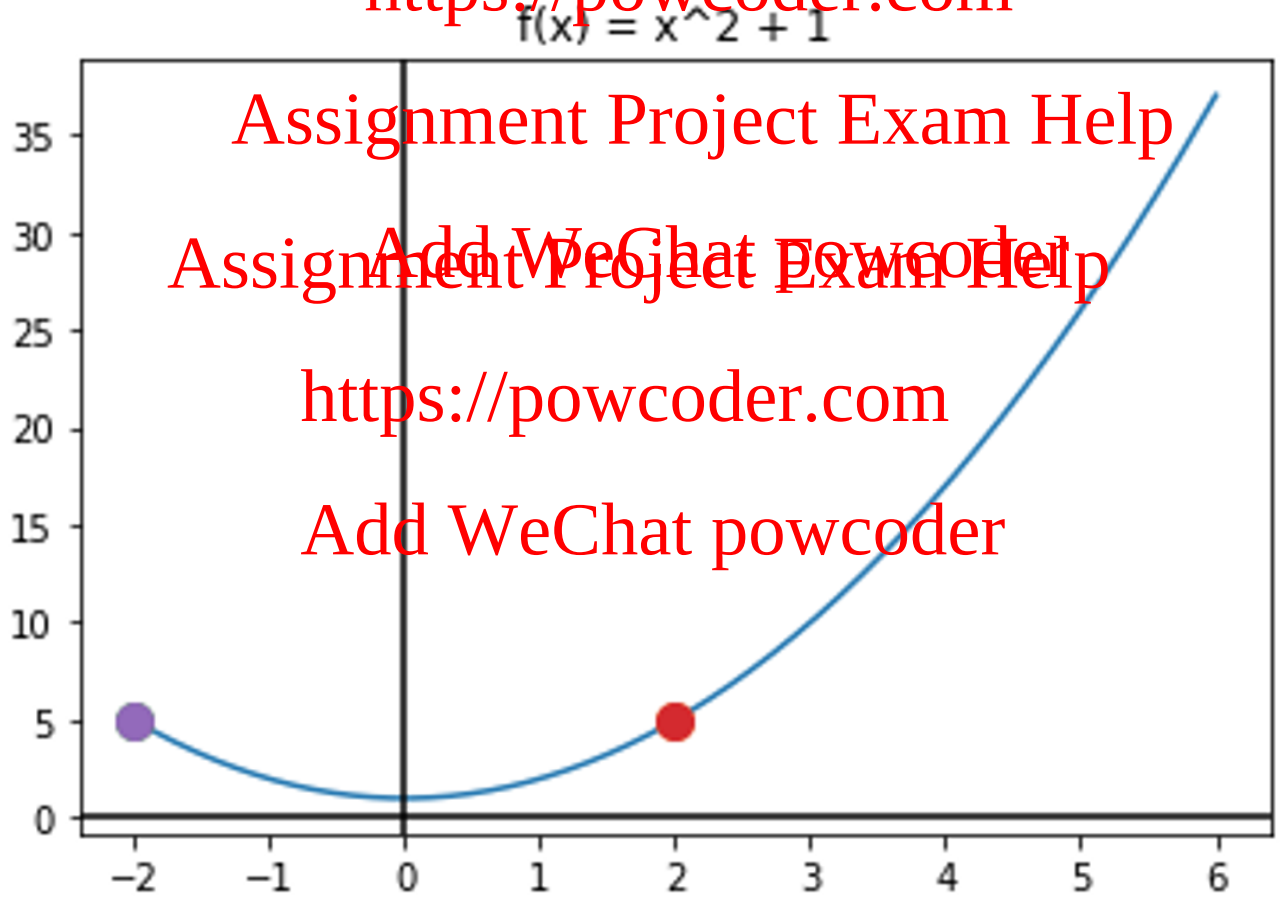
Let's still start at $x = -1$ and let's set the learning rate $\eta \triangleq 1$ instead and see what happens:

$$x = x - \Delta x = -1 - 1 \times (-2) = 1$$
$$f(x) = (1)^2 + 1 = 2$$
$$x = x - \Delta x = 1 - 1 \times (2) = -1$$
$$f(x) = (-1)^2 + 1 = 2$$
$$x = x - \Delta x = -1 - 1 \times (-2) = 1$$
$$f(x) = (1)^2 + 1 = 2$$

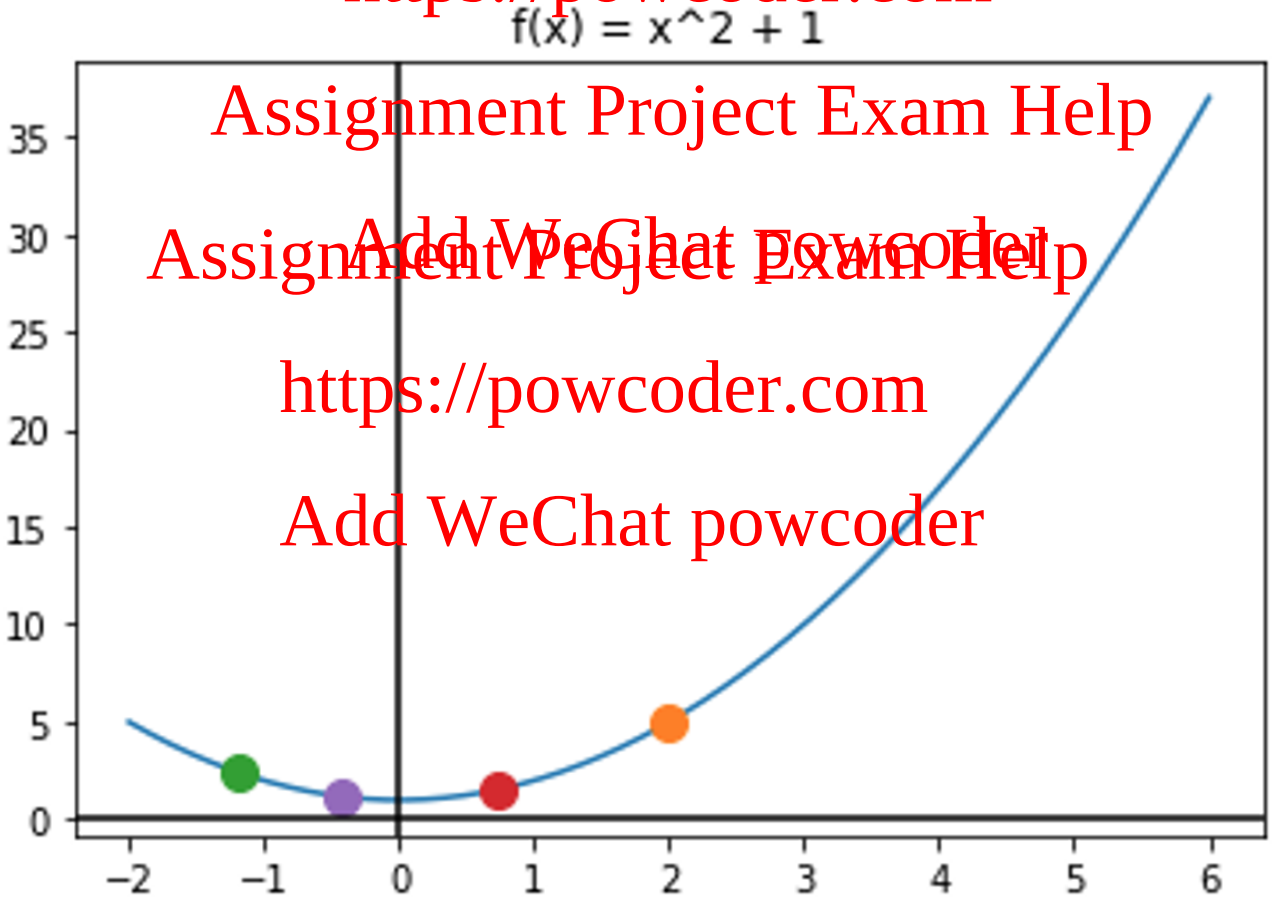
So the x will just swing back and forth without ever reaching the minimum.

Setting the right learning rate is thus very important. If set improperly, we'll never reach the minimum, or at least take much longer than necessary.

Trying to learn fast with a larger learning rate



Trying to learn fast with a larger learning rate



Derivative Rules

Common Derivatives:

$$\frac{d}{dx}(C) = 0 \qquad e.g., \frac{d}{dx}(91) = 0$$

$$\frac{d}{dx}(x) = 1 \qquad e.g., \frac{d}{dx}(4x) = 4$$

$$\frac{d}{dx}(x^n) = nx^{n-1} \qquad e.g., \frac{d}{dx}(x^4) = 4x^3$$

$$\frac{d}{dx}(a^x) = a^x \ln(a) \qquad e.g., \frac{d}{dx}(5^x) = 5^x \ln(5)$$

$$\frac{d}{dx}(e^x) = e^x$$

Note: ln: "Natural logarithm", logarithm to base of the mathematic constant e, where $e = 2.71882\cdots$

Derivative rules

More common derivatives:

$$\frac{d}{dx}(\ln(x)) = \frac{1}{x}, x > 0$$

$$\frac{d}{dx}(\ln(|x|)) = \frac{1}{x}, x \neq 0$$

$$\frac{d}{dx}(\log_a(x)) = \frac{1}{x \ln(a)}, x > 0$$

$$\frac{d}{dx}(\sin(x)) = \cos(x)$$

$$\frac{d}{dx}(\cos(x)) = -\sin(x)$$

$$\frac{d}{dx}(\tan(x)) = \sec^2(x)$$

Note: When $x <= 0$, $\ln(x)$ is unspecified. That is, you can't raise the constant e to any value to get a zero or a negative number.

Derivatives of functions

"The derivative of the function with respect to x"

$$\frac{d}{dx}(cf(x)) = c\frac{d}{dx}f(x)$$

$$\frac{d}{dx}(f(x) \pm g(x)) = \frac{d}{dx}f(x) \pm \frac{d}{dx}g(x)$$

$$\frac{d}{dx}(f(x)g(x)) = g(x)\frac{d}{dx}f(x) + f(x)\frac{d}{dx}g(x) \quad \text{(Product rule)}$$

$$\frac{d}{dx}\frac{f(x)}{g(x)} = \frac{g(x)\frac{d}{dx}f(x) - f(x)\frac{d}{dx}g(x)}{g^2} \quad \text{(Quotient rule)}$$

$$\frac{d}{dx}f(g(x)) = \frac{d}{dg(x)}f(g(x))\frac{d}{dx}g(x) \quad \text{(Chain rule)}$$

Breaking down the derivative of complex functions

Using these fundamental derivative rules, and particularly the chain rule, you can break down more complicated functions:

$$\frac{d}{dx}(f(x))^n = n(f(x)^{n-1})\frac{d}{dx}f(x)$$

$$\frac{d}{dx}e^{f(x)} = e^{f(x)}\frac{d}{dx}f(x)$$

$$\frac{d}{dx}\ln(f(x)) = \frac{1}{f(x)}\frac{d}{dx}f(x)$$

Partial Derivatives

- We don't normally deal with single variable functions in NLP. A typical NLP model (function) has tens of thousands or millions of variables (features). So we need to compute *partial derivatives*.
- Fortunately, compute partial derivatives is relatively simple. You just need to hold all other variables constant (treat them as constant), and take the derivative with respect to a given variable. $\frac{\partial}{\partial x} f(x, y)$, $\frac{\partial}{\partial y} f(x, y)$

More on partial derivatives

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

$$\frac{\partial}{\partial x} f(x,y) = 2x$$

$$\frac{\partial}{\partial y} f(x,y) = 2y$$

More on partial derivatives

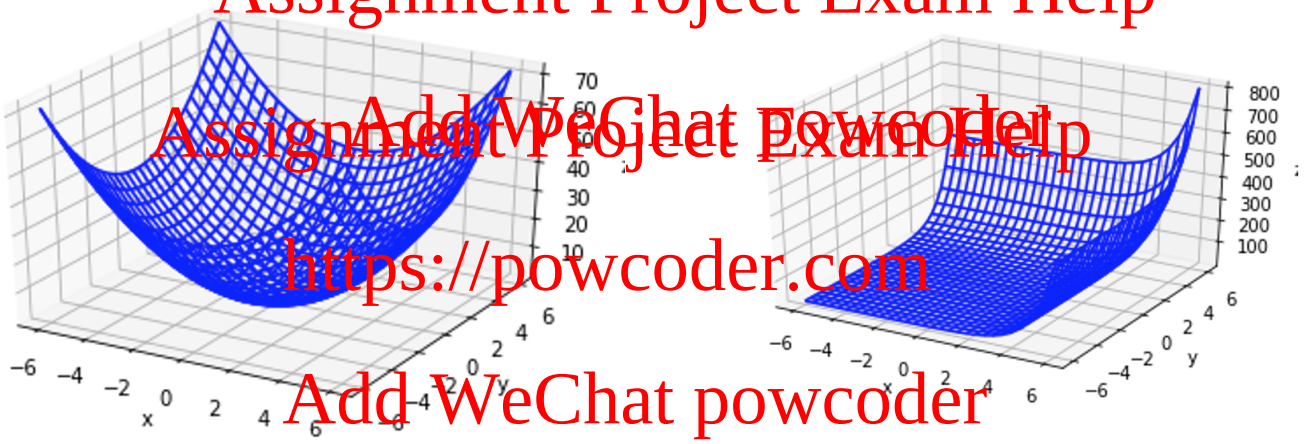
$$f(x,y) = \min(x,y) = \begin{cases} x & \text{if} \quad x < y \\ y & \text{if} \quad x > y \end{cases}$$

$$\frac{\partial}{\partial x} f(x,y) = \begin{cases} 1, & \text{if} \quad x < y \\ 0, & \text{if} \quad x > y \end{cases}$$

$$\frac{\partial}{\partial y} f(x,y) = \begin{cases} 0, & \text{if} \quad x < y \\ 1, & \text{if} \quad x > y \end{cases}$$

The function is not differentiable when $x = y$

Plot multi-variable functions



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

f(x,y) = e^x + e^y

Gradient

The gradient of a function ∇f is the set of partial derivatives of a function

$$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \cdots \\ \frac{\partial f}{\partial x_n} \end{bmatrix}$$

Properties of Logarithms

$$\log(xy) = \log(x) + \log(y) \qquad \ln(e^x) = x$$

$$\log(\frac{x}{y}) = \log(x) - \log(y) \qquad e^{\ln(x)} = x, \quad x > 0$$

$$\log(x^y) = y\log(x)$$

$$\log(\prod_i x_i) = \sum_i \log(x_i)$$

- It is common practice to map probabilities to logarithmic space to avoid *underflow* (when a value gets too close to zero for the computer to represent it).
  $\ln(0.0001) = -9.2103403\cdots$
- You can map the log values back to probabilities using the exponent. $e^{-9.2103403} = 0.0001$

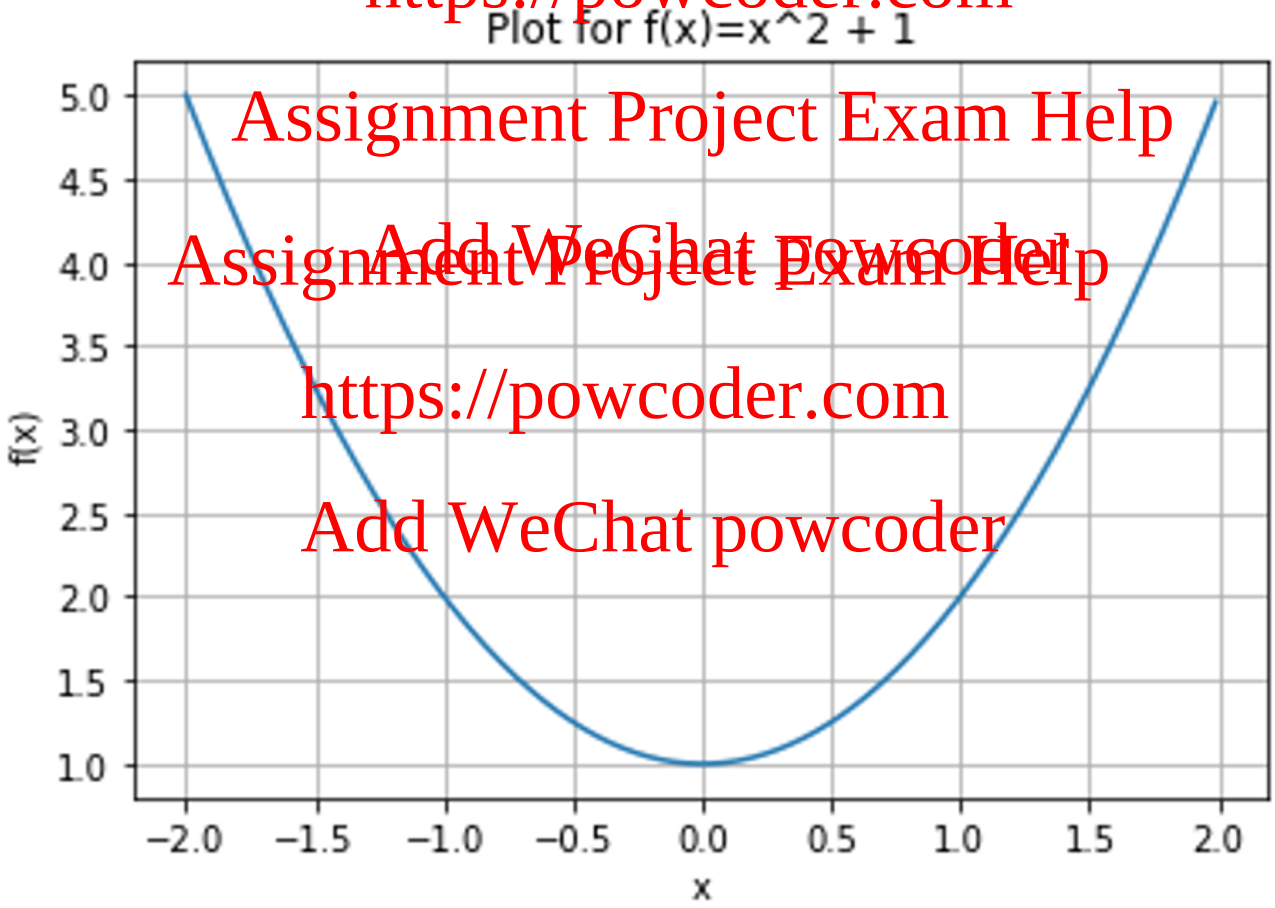
Convexity of functions

- Intuitively, a convex (conclave) function is a continuous function in which there is a single minimum (maximum).

- A mathematical definition: A convex function is a continuous function whose value at the midpoint of every interval in its domain does not exceed the arithmetic mean of its values at the ends of the interval.

- How to decide a function is convex: If $f(x)$ has a second derivative in $[a, b]$, then a necessary and sufficient condition for it to be convex on that interval is that the second derivative $f''(x) \geq 0$ for all x in $[a, b]$.

Example convex functions



Example non-convex functions