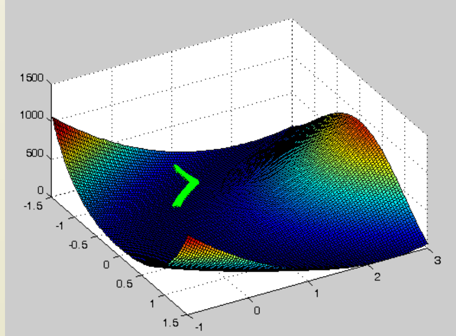
Logistic Regression

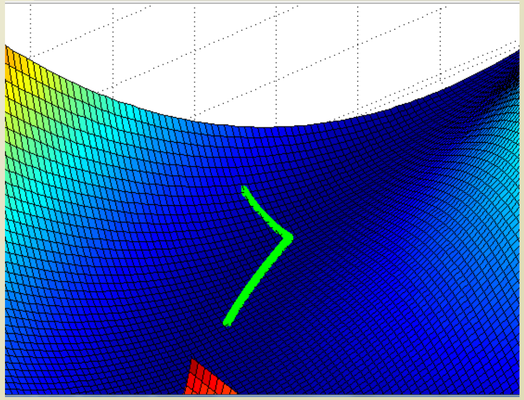
Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Gradient Descent/Ascent

- Gradient descent/ascent method is
  - Given a differentiable function of *f(x)* and an initial parameter of x_1
  - Iteratively moving the parameter to the lower/higher value of *f(x)*
  - By taking the direction of the negative/positive gradient of *f(x)*
- Why this works?
  - $f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + O(||x-a||^2)$ Useful Big-Oh Notation
    - Assume $a=x_1$ and $x=x_1+h\mathbf{u}$, $\mathbf{u}$ is the unit direction vector for the partial deriv.
    - $f(x_1+h\mathbf{u}) = f(x_1) + hf'(x_1)\mathbf{u} + h^2 O(1)$
    - $f(x_1+h\mathbf{u}) - f(x_1) \approx hf'(x_1)\mathbf{u}$ Always???
    - $\mathbf{u}^* = argmin_{\mathbf{u}}\{f(x_1+h\mathbf{u}) - f(x_1)\} = argmin_{\mathbf{u}} hf'(x_1)\mathbf{u} = -\frac{f'(x_1)}{|f'(x_1)|}$
    - $\because f(x_1+h\mathbf{u}) \le f(x_1), \vec{a}\cdot\vec{b} = |\vec{a}||\vec{b}|cos\alpha$ Gradient Descent
  - $x_{t+1} \leftarrow x_t + h\mathbf{u}^* = x_t - h\frac{f'(x_1)}{|f'(x_1)|}$
- Perfectly applicable to $\hat{\theta} = argmax_\theta \sum_{1\le i\le N} log(P(Y_i|X_i;\theta))$ Gradient Ascent
  - $f(\theta) = \sum_{1\le i\le N} log(P(Y_i|X_i;\theta))$
  - Setup an initial parameter of θ_1
  - Iteratively moving θ_t to the higher value of $f(\theta_t)$
  - By taking the direction of the ***positive*** gradient of $f(\theta_t)$

How Gradient Descent Works





- Example function: Rosenbrock function
  - $f(x_1, x_2) = (1 - x_1)^2 + 100(x_2 - x_1^2)^2$
  - $\frac{\partial}{\partial x_1} f(x_1, x_2) = -2(1 - x_1) - 400x_1(x_2 - x_1^2)$
  - $\frac{\partial}{\partial x_2} f(x_1, x_2) = 200(x_2 - x_1^2)$
- Assume the initial point
  - $\mathbf{x}^0 = (x_1^0, x_2^0) = (-1.3, 0.9)$

Global Minimum=0 at (1,1)

- Partial derivative vector at the point
  - $\boldsymbol{f}'(\mathbf{x}^0) = \left(\frac{\partial}{\partial x_1} f(x_1, x_2), \frac{\partial}{\partial x_2} f(x_1, x_2)\right) = (-415.4, -158)$
- Update the point with the negative partial derivative in a small scale, h=0.001
  - $\mathbf{x}^1 \leftarrow \mathbf{x}^0 - h \frac{f'(\mathbf{x}^0)}{|f'(\mathbf{x}^0)|}$
  - $\mathbf{x}^1 = \begin{pmatrix} -1.3 - 0.001 \times -415.4/444.4335, \\ 0.9 - 0.001 \times -158/444.4335 \end{pmatrix}$
  - $= (-1.2991, 0.9004)$
- Repeat the update until converges