Gradient Descent

Consider a function $E : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{w} = (w_1, w_2, \ldots, w_n) \mapsto E(\boldsymbol{w})$. The gradient ∇E of E is defined by

$$\nabla E := \left(\frac{\partial E}{\partial w_1}, \frac{\partial E}{\partial w_2}, \ldots, \frac{\partial E}{\partial w_n} \right).$$

Proposition

$E(\boldsymbol{w})$: differentiable in a nbhd of $\boldsymbol{w}$

The function $E(\boldsymbol{w})$ decreases fastest in the direction of $-\nabla E(\boldsymbol{w})$.

Proof: For a unit vector $\boldsymbol{u}$, the directional derivative $D_{\boldsymbol{u}}E$ is given by

$$D_{\boldsymbol{u}}E = \lim_{t \to 0} \frac{E(\boldsymbol{w}_0 + t\boldsymbol{u}) - E(\boldsymbol{w}_0)}{t} = g'(0),$$

where $g(t) = E(\boldsymbol{w}_0 + t\boldsymbol{u})$. Let $\boldsymbol{w} = \boldsymbol{w}_0 + t\boldsymbol{u}$. Using the chain rule,

$$g'(0) = \sum_{i=1}^{n} \frac{\partial E}{\partial w_i} \cdot \frac{dw_i}{dt} = \nabla E \cdot \boldsymbol{u}.$$

Furthermore,

$$D_{\boldsymbol{u}}E = \nabla E \cdot \boldsymbol{u} = |\nabla E|\,|\boldsymbol{u}|\cos\theta = |\nabla E|\cos\theta,$$

where θ is the angle between ∇E and $\boldsymbol{u}$. The minimum value of $D_{\boldsymbol{u}}E$ occurs when $\cos\theta$ is -1. $\square$

- Choose an initial point $\boldsymbol{w}_0$.
- Set

$$\boxed{\boldsymbol{w}_{k+1} = \boldsymbol{w}_k - \eta_k \nabla E(\boldsymbol{w}_k)}$$

where η_k is the step size or learning rate.

Usually, for some $\eta > 0$, we set $\eta_k = \eta$ for all k.

- From Proposition, we have

$$E(\boldsymbol{w}_k) \geq E(\boldsymbol{w}_{k+1}).$$

- Under some moderate conditions,

$$E(\boldsymbol{w}_k) \to \text{local minimum} \qquad \text{as } \; k \to \infty.$$

  In particular, this is true when E is convex or when ∇E is Lipschitz continuous.

- $\boldsymbol{w}_* = \lim_{k\to\infty} \boldsymbol{w}_k$

  We use $\boldsymbol{w}_k$ for a sufficiently large k as an approximation of $\boldsymbol{w}_*$.

  This method is called gradient descent.

  Caveat: Making a right choice of η is crucial.

Example

- Consider $E(\boldsymbol{w}) = E(w_1, w_2) = w_1^4 + w_2^4 - 16w_1w_2$.
  Then $\nabla E(\boldsymbol{w}) = [4w_1^3 - 16w_2, 4w_2^3 - 16w_1]$.
  Choose $\boldsymbol{w}_0 = (1, 1)$ and $\eta = 0.01$.
  $\boldsymbol{w}_{30} = (1.99995558586289, 1.99995558586289)$
  $E(\boldsymbol{w}_{30}) = -31.9999999368777$
- We see that $\boldsymbol{w}_k \to (2, 2)$ and $E(2, 2) = -32$.
- Indeed, when $\boldsymbol{w} = (2, 2)$, a local minimum of $E(\boldsymbol{w})$ is -32.
  Exercise: Find all the local minima of $E(\boldsymbol{w})$.

k	w1	w2	E(w1,w2)
1	1.12000000000000	1.12000000000000	-16.9233612800000
2	1.24300288000000	1.24300288000000	-19.9465014818312
3	1.36506297054983	1.36506297054983	-22.8698545020842
4	1.48172688079195	1.48172688079195	-25.4876645161458
5	1.58867706472624	1.58867706472624	-27.6422269714610
6	1.68247924276483	1.68247924276483	-29.2656452783487
7	1.76116971206054	1.76116971206054	-30.3861816086105
8	1.82445074094736	1.82445074094736	-31.0984991504577
9	1.87344669354831	1.87344669354831	-31.5194128485897
10	1.91018104795404	1.91018104795404	-31.7533053700606
11	1.93701591038872	1.93701591038872	-31.8770223901250
12	1.95622873443784	1.95622873443784	-31.9400248989010
13	1.96977907222858	1.96977907222858	-31.9712142030755
14	1.97923168007769	1.97923168007769	-31.9863406140263
15	1.98577438322011	1.98577438322011	-31.9935701975589
16	1.99027812738069	1.99027812738069	-31.9969902100773
17	1.99336647981957	1.99336647981957	-31.9985965516271
18	1.99547865709166	1.99547865709166	-31.9993473166738
19	1.99692058430943	1.99692058430943	-31.9996970174121
20	1.99790372262623	1.99790372262623	-31.9998595272283
21	1.99857347710339	1.99857347710339	-31.9999349274762
22	1.99902947615421	1.99902947615421	-31.9999698732955
23	1.99933981776146	1.99933981776146	-31.9999860577045
24	1.99955097148756	1.99955097148756	-31.9999935493971
25	1.99969461222478	1.99969461222478	-31.9999970160815
26	1.99979231393118	1.99979231393118	-31.9999986198712
27	1.99985876312152	1.99985876312152	-31.9999993617137
28	1.99990395413526	1.99990395413526	-31.9999997048203
29	1.99993468659806	1.99993468659806	-31.9999998634976
30	1.99995558586289	1.99995558586289	-31.9999999368777

Q: Can we do linear regression using gradient descent?

$$E = \|Y - XM\|^2, \qquad \nabla E = -2X^T(Y - XM)$$

- Define $\sigma(x) = \dfrac{e^x}{e^x + 1} = \dfrac{1}{1 + e^{-x}}$. It is called a sigmoid function.
- In logistic regression we will minimize the following error function

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

  where we write $\boldsymbol{w} = (w_1, w_2, \ldots, w_{k+1})$ and

  $y_n = \sigma(w_1 x_{n1} + w_2 x_{n2} + \cdots + w_k x_{nk} + w_{k+1})$.

- Compute the gradient $\nabla E(\boldsymbol{w})$.

- Crucial identity:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x))$$

Thus it is a solution to

$$\frac{dy}{dx} = y(1 - y),$$

which is called a logistic equation.

$$(\ln y)' = \frac{1}{y} y' = 1 - y.$$

$$(\ln(1 - y))' = \frac{1}{1 - y}(-y') = -y.$$

$$E(\boldsymbol{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\},$$

$$y_n = \sigma(w_1 x_{n1} + w_2 x_{n2} + \cdots + w_k x_{nk} + w_{k+1}).$$

- We have

$$\boxed{\frac{\partial E}{\partial w_j} = \sum_{n=1}^{N} (y_n - t_n) x_{nj}}.$$

- Assume $\nabla E = 0$.

  When $j = k + 1$, we have $x_{n,k+1} = 1$ for all n, and

$$\sum_{n=1}^{N} y_n = \sum_{n=1}^{N} t_n.$$

$$\sum_{n=1}^{N} y_n = \sum_{n=1}^{N} \frac{1}{1 + \exp(-(w_1 x_{n1} + w_2 x_{n2} + \cdots + w_k x_{nk} + w_{k+1}))}$$