Convex Loss Functions

Outline

- Softmax
- Convex Functions and MLE/MAP loss functions
- MLE/MAP ESTIMATION

Softmax

- findMin expects a 1d parameter vector w
- You can initialize w as follows

```
def fit(self,X, y):
    n, d = X.shape
    self.k = np.unique(y).size

    # Initialize w as one long vector
    self.w = np.zeros(d*self.k)

    # Expects 1D w vector
    utils.check_gradient(self, X, y)
    (self.w, f) = minimizers.findMin(...)

    # Reshape w to a 2D matrix
    self.w = np.reshape(self.w, (d, k))
```

Softmax

- findMin expects a 1d parameter vector *w*
- You can define the funObj method as follows

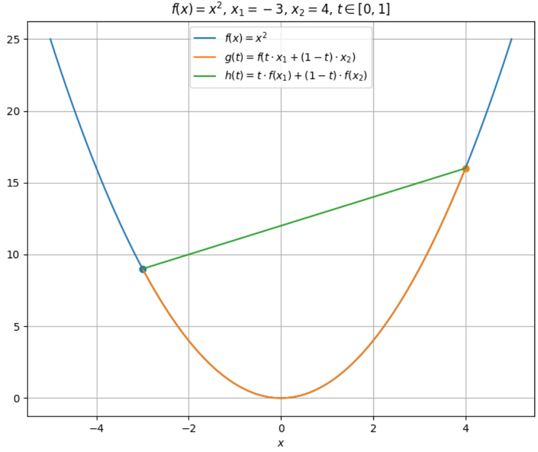
```
def funObj(self, w, X, y):
    n, d = X.shape
    # Reshape w to a 2D matrix
    W = np.reshape(w, (d, self.k))

    """ YOUR CODE HERE FOR COMPUTING 'f' and 'g' """

    # reshape gradient matrix to 1D vector
    return f, g.ravel()
```

- reshape and ravel return a view so no computation overhead

Definition of convexity

- A function f is convex if $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



1. Linear functions are convex

- $f(x) = Ax$ is a convex function
  - where A is some 2D matrix in $\mathbb{R}$
- **proof.**
  - A function f is convex if for $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0,1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

  - By definition, a linear function is:

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= A(tx_1 + (1-t)x_2) \\ &= tAx_1 + (1-t)Ax_2 \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{1}$$

  - Therefore, the linear function satisfies the convex inequality

2. Affine functions are convex

- $f(x) = Ax + b$ is convex where b is some vector in $\mathbb{R}$
- An Affine transformation is a linear transformation Ax plus translation b
  - All linear functions are affine functions but not vice versa
- **proof.**
  - A function f is convex if for $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

  - By definition, an affine function is:

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= A(tx_1 + (1-t)x_2) + b \\ &= tAx_1 + tb + (1-t)Ax_2 + (1-t)b \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{2}$$

  - Therefore, the affine function satisfies the convex inequality

3. Adding two convex functions results in a convex function

- $f(x) = h(x) + g(x)$ is a convex function
  - if $h(x)$ and $g(x)$ are convex
- **proof.**
  - A function f is convex if for $\forall x_1, x_2 \in \mathbb{R}; \forall t \in [0, 1]$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

  - Adding two convex functions:

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= h(tx_1 + (1-t)x_2) + g(tx_1 + (1-t)x_2) \\ &\leq th(x_1) + tg(x_1) + (1-t)h(x_2) + \\ &(1-t)g(x_2) \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{3}$$

4. Composition with an affine mapping

- $f(x) = g(Ax + b)$ is convex if g is convex
- **proof.**

$$\begin{aligned} f(tx_1 + (1-t)x_2) &= g(A(tx_1 + (1-t)x_2) + b) \\ &= g(t(Ax_1 + b) + (1-t)(Ax_2 + b)) \\ &\leq tg(Ax_1 + b) + (1-t)g(Ax_2 + b) \\ &= tf(x_1) + (1-t)f(x_2) \end{aligned} \tag{4}$$

- Therefore, knowing that $Ax + b$ is convex it is sufficient to show that $f(z)$ is convex by replacing $Ax + b$ with z.
  - might be helpful in the assignment.

5. Pointwise maximum

- The max of two convex functions is convex
- $f = \max(f_1, f_2)$ is convex
- **proof.**

$$
\begin{aligned}
f(tx_1 + (1-t)x_2) &= \max(f_1(tx_1 + (1-t)x_2), f_2(tx_1 + (1-t)x_2)) \\
&\leq \max(tf_1(x_1) + (1-t)f_1(x_2), tf_2(x_1) + (1-t)f_2(x_2)) \\
&\leq \max(tf_1(x_1), tf_2(x_1)) + \\
&\max((1-t)f_1(x_2), (1-t)f_2(x_2)) \\
&= tf(x_1) + (1-t)f(x_2)
\end{aligned}
\tag{5}
$$

5. Norms are convex functions

- For all norms $||x||_p = (\sum_{i=1}^{d} |x_i|^p)^{\frac{1}{p}}$ where $p \geq 1$ the following properties hold:
  - $||x|| \geq 0, \forall x \in R^d$
  - $||x|| = 0$ iff $x = 0$
  - $||ax|| = |a|||x||, \forall a \in R, x \in R^d$ (Homogeniety)
  - $||x_1 + x_2|| \leq ||x_1|| + ||x_2||, \forall x_1, x_2 \in R^d$ (Triangle inequality)
- **proof.** Norm functions are convex:

$$
\begin{aligned}
||tx_1 + (1-t)x_2|| &\leq ||tx_1|| + ||(1-t)x_2|| \quad \text{(Triangle Inequality)} \\
&= t||x_1|| + (1-t)||x_2|| \quad \text{(Homogeniety)}
\end{aligned}
\tag{6}
$$

6. Second-derivative test

- If the second derivative of a function $f(x)$ is positive $\forall x \in \mathbb{R}$ then f is convex
- **proof.**
- Using second order Taylor expansion, for some $\forall x_1, x_2 \in \mathbb{R}, \forall t \in [0,1]$:

$$f(x_2) = f(x_1) + \nabla f(x_1)^T(x_2 - x_1) + (x_2 - x_1)^T \nabla^2 f(x_1 + t(x_2 - x_1))(x_2 - x_1) \tag{7}$$

- Since $\nabla^2 f(x) > 0$

$$(x_2 - x_1)^T \nabla^2 f(x_1 + t(x_2 - x_1))(x_2 - x_1) \geq 0 \tag{8}$$

- Therefore,

$$f(x_2) \geq f(x_1) + \nabla f(x_1)^T(x_2 - x_1) \tag{9}$$

6. Second-derivative test proof

- Let $x_1 < x_2$ and $y = tx_1 + (1-t)x_2$, then

$$\begin{aligned} f(x_1) &\geq f(y) + \nabla f(y)^T (y - x_1) \\ f(x_2) &\geq f(y) + \nabla f(y)^T (y - x_2) \end{aligned} \tag{10}$$

- Multiply the first inequality by t and second by $(1-t)$ and add them to get,

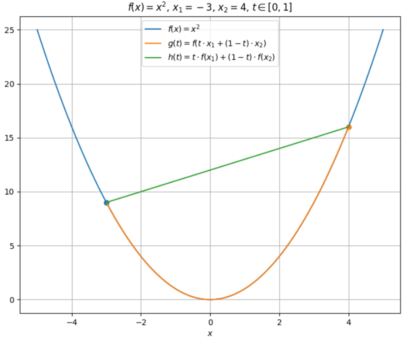
$$\begin{aligned} tf(x_1) + (1-t)f(x_2) &\geq tf(y) + (1-t)f(y) + \\ & t\nabla f(y)^T(y - x_1) + (1-t)\nabla f(y)^T(y - x_2) \\ & \Rightarrow tf(x_1) + (1-t)f(x_2) \geq f(y) + \\ & \nabla f(y)^T((t-1)x_1 + (1-t)x_2) + \nabla f(y)^T((t-1)x_2 + (1-t)x_1) \end{aligned} \tag{11}$$

- Therefore,

$$tf(x_1) + (1-t)f(x_2) \geq f(tx_1 + (1-t)x_2) \tag{12}$$

6. Second-derivative test

- Geometrically:
  - When $\nabla f(x)$ is negative, $f(x)$ decreases as x increases.
  - When $\nabla f(x)$ is positive, $f(x)$ increases as x increases.
  - Therefore, the minimum is at $x = a$ where the gradient switches sign.



MLE/MAP Estimation

- Maximum Likelihood (MLE)

$$\arg\max_w p(y|X, w)$$

  - Find w that makes y the highest probability given X, and w

- Maximum A Posteriori (MAP)

$$\arg\max_w p(w|X, y) \propto p(y|X, w) \cdot p(w)$$

  - Find w that maximizes its probability given X and y

Maximum Likelihood Estimation (MLE) example

- The normal distribution notation is $N(\mu, \sigma^2)$
- Given $y_i | x_i, w \sim N(w^T x_i, 1)$, which means:

$$p(y_i|x_i, w) = \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}) \tag{13}$$

- Therefore, maximizing $p(y|X, w)$ w.r.t w is equivalent to minimizing the unregularized least squares problem.

Maximum Likelihood Estimation (MLE) example

- Given

$$p(y|X, w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}) \tag{14}$$

$$\begin{aligned}
\arg\max_w p(y|X, w) &= \arg\max_w \prod_{i=1}^{N} \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}) \\
&= \arg\max_w \log(\frac{1}{\sqrt{2\pi}}) + \\
&\quad \sum_{i=1}^{N} \log(\exp(-\frac{(w^T x_i - y_i)^2}{2})) \quad \text{(log is monotonic)} \\
&= \arg\max_w - \sum_{i=1}^{N} \frac{(w^T x_i - y_i)^2}{2} \\
&= \arg\min_w \frac{1}{2} \cdot ||Xw - y||_2^2 \quad \text{(negate both sides)} \\
&= \arg\min_w ||Xw - y||_2^2 \quad \text{(does not change solution)}
\end{aligned}$$

Maximum A Posteriori (MAP) example

- Given

$$p(y|X,w) = \prod_{i=1}^{N} \frac{1}{\sqrt{2 \cdot 1 \cdot \pi}} \exp(-\frac{(w^T x_i - y_i)^2}{2 \cdot 1}) \quad y_i|x_i, w \sim N(w^T x_i, 1)$$

$$p(w) = \prod_{j=1}^{d} \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}}) \quad w_j \sim N(0, \lambda^{-1}) \tag{16}$$

$$\begin{aligned}
\arg\max_w p(w|X,y) &= \arg\max_w p(y|X,w) \cdot p(w) \\
&= \arg\max_w \log(p(y|X,w)) + \log(\prod_{j=1}^{d} \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}})) \\
&= \arg\max_w \log(p(y|X,w)) + \sum_{j=1}^{d} \log(\exp(-\frac{\lambda}{2} w_j^2))
\end{aligned} \tag{17}$$

Maximum A Posteriori (MAP) example

$$
\begin{aligned}
\arg\max_w p(w|X,y) &= \arg\max_w p(y|X,w) \cdot p(w) \\
&= \arg\max_w \log(p(y|X,w)) + \log(\prod_{j=1}^{d} \frac{1}{\sqrt{2 \cdot \lambda^{-1} \cdot \pi}} \exp(-\frac{(w_j - 0)^2}{2 \cdot \lambda^{-1}})) \\
&= \arg\max_w \log(p(y|X,w)) + \sum_{i=1}^{d} \log(\exp(-\frac{\lambda}{2} w_j^2)) \\
&= \arg\min_w -\log(p(y|X,w)) + \sum_{i=1}^{d} \frac{\lambda}{2} w_j^2 \quad \text{(negate both sides)} \\
&= \arg\min_w \frac{1}{2}||Xw - y||^2 + \frac{\lambda}{2}||w||^2
\end{aligned}
\tag{18}
$$

Maximum A Posteriori (MAP) Assignment

- For a given objective function, follow the following steps:
  - Simplify using operations such as applying the log and negative sign
  - Show the new loss function.
- Given

$$
\begin{aligned}
p(y|X, w) &= \ ? \\
p(w) &= \ ?
\end{aligned}
\tag{19}
$$

$$
\begin{aligned}
\arg\max_{w} P(w|X, y) &= \arg\max_{w} p(y|X, w) \cdot p(w) \\
&= \arg\min_{w} \quad ?
\end{aligned}
\tag{20}
$$