Problem 1 – Ali-Silvey divergences are convex

If $\phi$ convex, then $t\phi(x/t)$ is also convex for $t > 0$; if we consider $p, q$ non-negative vectors in $\mathbb{R}^{\mathbb{R}^{|\Omega|}}$, then $d(p, q)$ is a sum of convex functions, so it is convex. If we in addition impose that $p, q$ be normalized, we are restricting $d$ to an affine space, and this restriction is convex as well.

Problem 2 – Stochastic gradient for regularized regression

2.1

$$\nabla f = R\beta - \frac{2}{N} \sum_i (y^i - \beta^T x^i)x^i \quad (1)$$

$$= (RI + \frac{2}{N} \sum_i x^i(x^i)^T)\beta - \frac{2}{N} \sum_i y^ix^i \quad (2)$$

$$\nabla^2 f = RI + \frac{2}{N} \sum_i x^i(x^i)^T \quad (3)$$

2.2 Stochastic Gradient Descent

Initialize $\beta^1 = 0$ for $k = 1 : K$

1. sample $(x^k, y^k)$ uniformly from $D$

2. $d^k \leftarrow R\beta^k + 2x^k[(x^k)^T \beta^k - y^k]$

3. $\beta^{k+1} \leftarrow \beta^k (1 - \frac{cR}{M}) + \frac{2c}{M}[y^k - (\beta^k)^T x^k]x^k$

4. update $\bar{\beta}$

Output $\bar{\beta}$

2.3 Each example is sampled with equal probability $1/N$. The expectation of $d$ is

$$\frac{1}{N} \sum_i d(x^i, y^i) = R\beta - \frac{2}{N} \sum_i (y^i - \beta^T x^i)x^i = \nabla f \quad (4)$$
Problem 3 – Discrete AdaBoost

<table>
<thead>
<tr>
<th>data $x^i$</th>
<th>$x^1$</th>
<th>$x^2$</th>
<th>$x^3$</th>
<th>$x^4$</th>
<th>$x^5$</th>
<th>$x^6$</th>
<th>$x^7$</th>
<th>$x^8$</th>
<th>$x^9$</th>
<th>$x^{10}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>true labels $y^i$</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>−</td>
<td>+</td>
<td>−</td>
</tr>
<tr>
<td>$w^1_i$</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

$b^1(x^i)$ | + | + | + | + | − | − | − | − | − | − |

$[y^i b^1(x^i)]$ | − | − | − | − | − | − | − | − | − | − |

$r^1 = \sum_i w^1_i y^i b^1(x^i) = 0.8$
$\beta^1 = \frac{1}{2} \ln \frac{11}{8} = \ln 3$
$e^{\beta^1} = 3$

$w^2_i$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $3$ | $\frac{1}{3}$ | $\frac{1}{3}$ | $\frac{1}{3}$ |

$b^2(x^i)$ | + | + | + | + | + | + | + | + | − | − |

$[y^i b^2(x^i)]$ | + | + | + | + | + | + | + | + | − | − |

$r^2 = \sum_i w^2_i y^i b^2(x^i) = 2/3$
$\beta^2 = \frac{1}{2} \ln \frac{5/3}{1/3} = \frac{1}{2} \ln \frac{5}{2}$
$e^{\beta^2} = \sqrt{\frac{5}{2}}$

Problem 4 – Linear regression with uniform noise

4.1

$$p(y|x, \theta, t) = \begin{cases} \frac{1}{2t} & \text{if } |x^T \theta - y| \leq t \\ 0 & \text{otherwise} \end{cases}$$ (5)

4.2

$$p(y^{1:N}|x^{1:N}, \theta, t) = \begin{cases} \frac{1}{2t^N} & \text{if } |(x^i)^T \theta - y^i| \leq t \text{ for all } i = 1 : N \\ 0 & \text{otherwise} \end{cases}$$ (6)

Maximizing $1/(2t)^N$ is the same as minimizing $t$. Hence, the Maximum likelihood estimation problem for $\theta, t$ can be written as

$$\min_{x,t} \quad t$$ (7)

s.t \hspace{1cm} |(x^i)^T \theta - y^i| \leq t \text{ for all } i = 1 : N \quad (8)

$$t \geq 0 \quad (9)$$

or, equivalently,

$$\min_{x,t} \quad t$$ (10)

s.t \hspace{1cm} -t \leq (x^i)^T \theta - y^i \leq t \text{ for all } i = 1 : N \quad (11)

$$t \geq 0 \quad (12)$$

4.3 Denote the dual parameters associated with the constraints in the first row by $\lambda^i_-$ respectively $\lambda^i_+$, and the dual variable associated with the last
constraint by $\tau$. The Lagrangean is
\[
L(\theta, t, \lambda^\pm, \tau) = t + \sum_i [\lambda_i^+ ((x^i)^T \theta - y^i - t) - \lambda_i^- ((x^i)^T \theta - y^i + t)] - \tau t \quad (13)
\]

4.4 We minimize $L$ w.r.t. the primal variables.

\[
\frac{\partial L}{\partial \theta_i} = (\lambda_i^+ - \lambda_i^-) x^i 
\]
\[
\frac{\partial L}{\partial t} = 1 - \sum_i (\lambda_i^+ + \lambda_i^-) - \tau \quad (15)
\]

The linear constraints cancel the linear terms in $\theta$ and $t$ from the Lagrangean, and the dual function is obtained as
\[
g(\lambda^\pm, \tau) = - (\lambda^+ - \lambda^-)^T y \quad (16)
\]

where $\lambda^\pm, y$ denote respectively the $N$ dimensional vectors containing $\lambda_i^\pm$ and $y^i$. The dual problem is

\[
\begin{align*}
\max_{\lambda^\pm, \tau} & \quad - (\lambda^+ - \lambda^-)^T y \\
\text{s.t.} & \quad X^T (\lambda^+ - \lambda^-) = 0 \\
 & \quad 1^T (\lambda^+ - \lambda^-) = 1 - \tau \\
 & \quad \lambda^\pm, \tau \geq 0
\end{align*} 
\]

(17) (18) (19) (20)

The variable $\tau$ can be eliminated, obtaining

\[
\begin{align*}
\max_{\lambda^\pm} & \quad - (\lambda^+ - \lambda^-)^T y \\
\text{s.t.} & \quad X^T (\lambda^+ - \lambda^-) = 0 \\
 & \quad 1^T (\lambda^+ - \lambda^-) \leq 1 \\
 & \quad \lambda^\pm \geq 0
\end{align*} \quad (21) \quad (22) \quad (23) \quad (24)
\]

4.5 Under the assumption, the optimal $t$ must be greater than 0. Therefore, the constraints $\theta^T x^i - y \leq t$ and $\theta^T x^i - y \geq -t$ cannot be tight simultaneously. At least one of them must be slack, and the corresponding $\lambda_i$ will consequently be 0. Therefore the desired result follows.

Note also: Under the assumption in 4.5, $\lambda_i^+, \lambda_i^-$ can be considered the positive and negative part of a real number $\lambda_i = \lambda_i^+ - \lambda_i^-$. Then $|\lambda_i| = \lambda_i^+ + \lambda_i^-$. It follows that the dual problem can be transformed into
\[
\begin{align*}
\min_{\lambda} & \quad y^T \lambda \\
\text{s.t.} & \quad X^T \lambda = 0, \quad ||\lambda||_1 \leq 1.
\end{align*} \quad (25)
\]

3
Problem 5 – Coordinate descent versus stochastic gradient

\[ Q = \begin{bmatrix} q_{11} & q_{12} \\ q_{12} & q_{22} \end{bmatrix} \]

with \( q_{11} \neq q_{22}, q_{12} \neq 0 \). We run the following algorithm

**Stochastic Coordinate Descent**

**Initialize** at \( x^0, y^0 \) (with \( x^0, y^0 \neq 0 \))

for \( k = 1, 2, \ldots \)

1. Chose randomly coordinate \( x \) or coordinate \( y \)
2. Find the minimum \( x^{\text{new}} \) (or \( y^{\text{new}} \)) along the chosen direction
3. \( (x^{k+1}, y^{k+1}) \leftarrow (x^k, y^{\text{new}}) \) if minimization over \( y \) or \( (x^{k+1}, y^{k+1}) \leftarrow (x^{\text{new}}, y^k) \) if minimization over \( x \)

5.1

\[
\frac{\partial f}{\partial x} = 2(q_{11}x + q_{12}y) \Rightarrow x^{\text{new}} = \frac{-q_{12}y}{q_{11}} \tag{26}
\]

\[
\frac{\partial f}{\partial y} = 2(q_{12}x + q_{22}y) \Rightarrow y^{\text{new}} = \frac{-q_{12}x}{q_{22}} \tag{27}
\]

5.2 One solution is to compare the steps in each direction taken by stochastic gradient and by **Stochastic Coordinate Descent**. If they are proportional, irrespective of direction, then there could be a step size for stochastic gradient to achieve the same steps as **Stochastic Coordinate Descent**. As it turns out

\[
\frac{x^{\text{new}} - x}{\frac{\partial f}{\partial x}} = \frac{1}{2q_{11}} \tag{28}
\]

\[
\frac{y^{\text{new}} - y}{\frac{\partial f}{\partial y}} = \frac{1}{2q_{22}} \tag{29}
\]

These ratio are constant in each direction, but different. Thus, **Stochastic Coordinate Descent** is NOT a stochastic gradient algorithm.

*Note that if the stochastic coordinate descent took decreasing step sizes then it would be identical to stochastic gradient.*