Combinatorial Optimization in Computer Vision
Lecture: F. R. Schmidt and C. Domokos
Exercises: T. Möllenhoff and T. Windheuser
Winter Semester 2015/2016

Computer Vision Group
Institut für Informatik
Technische Universität München

Weekly Exercises 12

Room: 02.09.023
Tuesday, 02.02.2016, 14:15-15:45
Submission deadline: Tuesday, 02.02.2016, 11:15 , Room 02.09.023

1 Parameter Learning (15 Points)

Exercise 1 (2 Points)**.** Let $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\}$ be a set of given identically independently distributed (i.i.d.) observations. By assuming that w is a random vector with prior distribution $p(w)$, show that the following posterior distribution can be written as:

$$p(w|\mathcal{D}) = p(w) \prod_{n=1}^{N} \frac{p(y_n|x_n, w)}{p(y_n|x_n)}.$$

Solution. Assuming that the observations are i.i.d.:

$$\begin{aligned} p(w|\mathcal{D}) &= \frac{p(y_1, \ldots, y_N, x_1, \ldots, x_N, w)}{p(y_1, \ldots, y_N, x_1, \ldots, x_N)} = \frac{p(y_1, \ldots, y_N|x_1, \ldots, x_N, w)p(w)p(x_1, \ldots, x_N)}{p(y_1, \ldots, y_N|x_1, \ldots, x_N)p(x_1, \ldots, x_N)} \\ &= \frac{p(y_1, \ldots, y_N|x_1, \ldots, x_N, w)p(w)}{p(y_1, \ldots, y_N|x_1, \ldots, x_N)} = p(w) \prod_{n=1}^{N} \frac{p(y_n|x_n, w)}{p(y_n|x_n)} \end{aligned}$$

Exercise 2 (2 Points)**.** Calculate the expected loss $\mathbb{E}_{y \sim d(y|x)} [\Delta_H(y, f(x))]$ of the Hamming loss:

$$\Delta_H(y, y') = \frac{1}{|V|} \sum_{i \in V} [\![y_i \neq y'_i]\!],$$

where $d(y|x)$ denotes the true data distribution.

Solution.

$$\begin{aligned}
\mathbb{E}_{y\sim d(y|x)}\left[\Delta_H(y, f(x))\right] &= \sum_{y\in Y} d(y|x)\Delta_H(y, f(x)) \approx \sum_{y\in Y} p(y|x, w)\Delta_H(y, f(x)) \\
&= \sum_{y\in Y} p(y|x, w)\frac{1}{|V|}\sum_{i\in V}[\![y_i \neq f(x)_i]\!] \\
&= \sum_{y\in Y} p(y|x, w)\frac{1}{|V|}\left(\sum_{i\in V} 1 - [\![y_i = f(x)_i]\!]\right) \\
&= \sum_{y\in Y} p(y|x, w) - \sum_{y\in Y} p(y|x, w)\frac{1}{|V|}\left(\sum_{i\in V}[\![y_i = f(x)_i]\!]\right) \\
&= 1 - \frac{1}{|V|}\sum_{i\in V} p(y_i = f(x)_i|x, w).
\end{aligned}$$

Exercise 3 (6 Points)**.** Consider the negative regularized conditional log-likelihood (cf. Lecture 22):

$$\mathcal{L}(w) = \lambda \|w\|^2 + \sum_{n=1}^{N}\langle w, \varphi(x^n, y^n)\rangle + \sum_{n=1}^{N}\log Z(x^n, w). \tag{1}$$

It has been shown in the lecture that the gradient of (1) w.r.t. w is given as

$$\nabla_w\mathcal{L}(w) = 2\lambda w + \sum_{n=1}^{N}\left[\varphi(x^n, y^n) - \mathbb{E}_{y\sim p(y|x^n)}[\varphi(x^n, y)]\right].$$

Show that the Hessian of (1) is given as

$$\begin{aligned}
H_w\mathcal{L}(w) = 2\lambda I + \sum_{n=1}^{N} &\mathbb{E}_{y\sim p(y|x^n)}[\varphi(x^n, y)\varphi(x^n, y)^\mathsf{T}] - \\
&\mathbb{E}_{y\sim p(y|x^n)}[\varphi(x^n, y)]\mathbb{E}_{y\sim p(y|x^n)}[\varphi(x^n, y)]^\mathsf{T},
\end{aligned}$$

and argue that it is positive definite for $\lambda > 0$.

Solution. The gradient is a function $g : \mathbb{R}^n \to \mathbb{R}^n$ given as the following

$$g(w) = 2\lambda w + \sum_{i=1}^{N}\varphi(x^n, y^n) - \sum_{i=1}^{N}\sum_{y\in Y}\frac{\exp(-\langle w, \varphi(x^n, y)\rangle)}{\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)}\varphi(x^n, y).$$

The Hessian matrix is now defined as

$$H_w = \begin{pmatrix}
\frac{\partial g_1}{\partial w_1} & \frac{\partial g_2}{\partial w_1} & \cdots & \frac{\partial g_n}{\partial w_1} \\
\frac{\partial g_1}{\partial w_2} & \frac{\partial g_2}{\partial w_2} & \cdots & \frac{\partial g_n}{\partial w_2} \\
\vdots & \vdots & \vdots & \vdots \\
\frac{\partial g_1}{\partial w_n} & \frac{\partial g_2}{\partial w_n} & \cdots & \frac{\partial g_n}{\partial w_n}
\end{pmatrix},$$

where $g_i : \mathbb{R}^n \to \mathbb{R}$ denotes the i-th component of $g : \mathbb{R}^n \to \mathbb{R}^n$. We compute the partial derivative of g_i w.r.t. w_j as the following:

$$\begin{aligned}\frac{\partial g_i}{\partial w_j}(w) &= \frac{\partial g_i}{\partial w_j}(2\lambda w_i) - \sum_{i=1}^{N}\sum_{y\in Y}\frac{\partial g_i}{\partial w_j}\left(\frac{\exp(-\langle w, \varphi(x^n, y)\rangle)}{\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)}\varphi_i(x^n, y)\right)\\ &= 2\lambda[\![i = j]\!] - \sum_{i=1}^{N}\sum_{y\in Y}\dots\end{aligned}$$

For the expression "…" inside the sums, we apply the quotient rule:

$$\frac{\partial g_i}{\partial w_j}\left(\frac{\exp(-\langle w, \varphi(x^n, y)\rangle)}{\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)}\varphi_i(x^n, y)\right) =$$

$$\frac{\left(\frac{\partial g_i}{\partial w_j}\exp(-\langle w, \varphi(x^n, y)\rangle)\varphi_i(x^n, y)\right)\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)}{\left(\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)\right)^2} -$$

$$\frac{\exp(-\langle w, \varphi(x^n, y)\rangle)\varphi_i(x^n, y)\left(\sum_{y'\in Y}\frac{\partial g_i}{\partial w_j}\exp(-\langle w, \varphi(x^n, y')\rangle)\right)}{\left(\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)\right)^2}$$

Now using

$$\frac{\partial g_i}{\partial w_j}\left(\exp(-\langle w, \varphi(x^n, y)\rangle)\right) = -\exp(-\langle w, \varphi(x^n, y)\rangle)\varphi_j(x^n, y)$$

the above further simplifies to

$$-\frac{(\exp(-\langle w, \varphi(x^n, y)\rangle)\varphi_j(x^n, y)\varphi_i(x^n, y))}{\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)} -$$

$$\frac{\exp(-\langle w, \varphi(x^n, y)\rangle)\varphi_i(x^n, y)\left(\sum_{y'\in Y} -\exp(-\langle w, \varphi(x^n, y')\rangle)\varphi_j(x^n, y')\right)}{\left(\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)\right)^2} =$$

$$-p(y|x^n, w)\varphi_j(x^n, y)\varphi_i(x^n, y) - p(y|x^n, w)\sum_{y''\in Y}\frac{-\exp(-\langle w, \varphi(x^n, y'')\rangle)}{\sum_{y'\in Y}\exp(-\langle w, \varphi(x^n, y')\rangle)}\varphi_i(x^n, y)\varphi_j(x^n, y'') =$$

$$-p(y|x^n, w)\varphi_j(x^n, y)\varphi_i(x^n, y) + p(y|x^n, w)\varphi_i(x^n, y)\sum_{y'\in Y}p(y'|x^n, w)\varphi_j(x^n, y') =$$

$$-p(y|x^n, w)\varphi_j(x^n, y)\varphi_i(x^n, y) + p(y|x^n, w)\varphi_i(x^n, y)\mathbb{E}_{y'\sim p(y'|x,w)}[\varphi_j(x^n, y')].$$

Substituing back this yields:

$$
\begin{aligned}
\frac{\partial g_i}{\partial w_j}(w) &= 2\lambda[\![i=j]\!] - \sum_{i=1}^{N}\sum_{y\in Y} -p(y|x^n,w)\left(\varphi_j(x^n,y)\varphi_i(x^n,y) + \varphi_i(x^n,y)\mathbb{E}_{y'\sim p(y'|x,w)}[\varphi_j(x^n,y')]\right) \\
&= 2\lambda[\![i=j]\!] - \sum_{i=1}^{N} -\mathbb{E}_{y\sim p(y|x^n,w)}\left[\varphi_j(x^n,y)\varphi_i(x^n,y) + \varphi_i(x^n,y)\mathbb{E}_{y'\sim p(y'|x,w)}[\varphi_j(x^n,y')]\right] \\
&= 2\lambda[\![i=j]\!] - \sum_{i=1}^{N} -\mathbb{E}_{y\sim p(y|x^n,w)}\left[\varphi_j(x^n,y)\varphi_i(x^n,y)\right] + \\
&\qquad \mathbb{E}_{y\sim p(y|x^n,w)}\left[\mathbb{E}_{y'\sim p(y'|x,w)}[\varphi_i(x^n,y)\varphi_j(x^n,y')]\right] \\
&= 2\lambda[\![i=j]\!] - \sum_{i=1}^{N} -\mathbb{E}_{y\sim p(y|x^n,w)}\left[\varphi_j(x^n,y)\varphi_i(x^n,y)\right] + \\
&\qquad \mathbb{E}_{y\sim p(y|x^n,w)}\left[\varphi_i(x^n,y)\right]\mathbb{E}_{y\sim p(y'|x,w)}\left[\varphi_j(x^n,y)\right].
\end{aligned}
$$

In matrix-vector form this can be written as:

$$
\begin{aligned}
H_w(w) &= 2\lambda I + \sum_{i=1}^{N} \mathbb{E}_{y\sim p(y|x^n,w)}[\varphi(x^n,y)\varphi(x^n,y)^\mathsf{T}] - \mathbb{E}_{y\sim p(y|x^n,w)}[\varphi(x^n,y)]\mathbb{E}_{y\sim p(y|x^n,w)}[\varphi(x^n,y)]^\mathsf{T} \\
&= 2\lambda I + \sum_{i=1}^{N} \operatorname{cov}\left[\varphi(x^n,y),\varphi(x^n,y)\right].
\end{aligned}
$$

Furthermore covariance matrices $\operatorname{cov}[X,X] = \mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T]$ are always positive semi-definite:

$$
\begin{aligned}
u^T\mathbb{E}[(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T]u &= \mathbb{E}[u^T(X-\mathbb{E}[X])(X-\mathbb{E}[X])^T u] \\
&= \mathbb{E}[(u^T(X-\mathbb{E}[X]))^2] \geq 0.
\end{aligned}
$$

Adding $2\lambda I$ results in the Hessian to be positive definite.

Exercise 4 (2 Points)**.** Let $\xi \sim \mathcal{U}(-1,1)$ be a continuous random variable which follows a uniform distribution on the interval $[-1,1]$. Calculate the cumulative distribution function of the variable $\nu \sim \xi^2$.

Solution. The CDF of ξ is given as

$$
F(x) = \begin{cases} 0, & \text{if } x \leq -1, \\ \frac{x+1}{2}, & \text{if } -1 \leq x \leq 1, \\ 1, & \text{if } x \geq 1. \end{cases}
$$

To compute the cumulative distribution G of $\nu = \xi^2$ in terms of F note that:

$$
G(y) = P(\nu \leq y) = P(\xi^2 \leq y) = P(-\sqrt{y} \leq \xi \leq \sqrt{y}) = F(\sqrt{y}) - F(-\sqrt{y}).
$$

We compute

$$F(\sqrt{y}) = \begin{cases} \frac{\sqrt{y}+1}{2}, & \text{if } 0 \leq y \leq 1, \\ 1, & \text{if } y > 1, \end{cases}$$

and

$$F(-\sqrt{y}) = \begin{cases} \frac{1-\sqrt{y}}{2}, & \text{if } 0 \leq y \leq 1, \\ 0, & \text{if } y \geq 1, \end{cases}$$

Hence we have for $G(y) = F(\sqrt{y}) - F(-\sqrt{y})$:

$$G(y) = \begin{cases} 0, & \text{if } y < 0, \\ \sqrt{y}, & \text{if } 0 \leq y \leq 1, \\ 1, & \text{if } y > 1. \end{cases}$$

Exercise 5 (3 Points)**.** Compute the *subdifferential* at a point $x \in \mathbb{R}^n$

$$\partial f(x) = \{w \in \mathbb{R}^n \mid f(x) + \langle w, y - x \rangle \leq f(y), \forall y \in \mathbb{R}^n\},$$

of the following convex functions $f : \mathbb{R}^n \to \mathbb{R}$:

a) $f(x) = \langle c, x \rangle$, where $c \in \mathbb{R}^n$ is a constant

b) $f(x) = \|x\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

c) $f(x) = \|x\|_1 = \sum_{i=1}^n |x_i|$

Solution.

a) By definition of $\partial f(x)$:

$$\begin{aligned} &\langle c, x \rangle + \langle w, y - x \rangle \leq \langle c, y \rangle, \quad \forall y \in \mathbb{R}^n \\ &\Leftrightarrow \langle c - w, x - y \rangle \leq 0, \quad \forall y \in \mathbb{R}^n \\ &\Leftrightarrow w = c. \end{aligned}$$

Hence $\partial f(x) = \{c\}$. Note that if f is differentiable at $x \in \mathbb{R}^n$ we have that $\partial f(x) = \{\nabla f(x)\}$.

b) If $x = 0$:

$$\begin{aligned} &\langle w, y \rangle - \|y\|_2 \leq 0, \quad \forall y \in \mathbb{R}^n \\ &\langle w, y \rangle - \|y\|_2 \leq \|w\|_2 \|y\|_2 - \|y\|_2 = (\|w\|_2 - 1) \|y\|_2 \leq 0. \end{aligned}$$

Hence $\partial f(0) = \{w \in \mathbb{R}^n \mid \|w\|_2 \leq 1\}$.
For $x \neq 0$, $f(x)$ is differentiable, hence:

$$\partial f(x) = \left\{ \frac{x}{\|x\|_2} \right\}.$$

c) Again for $x \neq 0$, $f(x)$ is differentiable with slope $+1$ if $x_i > 0$ and -1 if $x_i < 0$:

$$\partial f(x) = \{\mathrm{sgn}(x)\},$$

where $\mathrm{sgn}(x)$ is the vectorial sign function.
For $x = 0$ we have:

$$\begin{aligned} &\sum_i w_i y_i - \sum |y_i| \leq 0, \quad \forall y \in \mathbb{R}^n, \\ &\Leftrightarrow |w_i| \leq 1, \quad \forall 1 \leq i \leq n. \end{aligned}$$

Thus $\partial f(x) = \{w \in \mathbb{R}^n \mid \|w\|_\infty \leq 1\}$.