Module: M2. Optimization and inference techniques for Computer Vision **Final exam**
Date: December 1rst, 2016
Teachers: Juan Fco Garamendi, Coloma Ballester, Oriol Ramos, Joan Serrat **Time: 2h30min**

- Books, lecture notes, calculators, phones, etc. are not allowed.
- All sheets of paper should have your name.
- Answer each problem in a separate sheet of paper.
- All results should be demonstrated or justified.

Problem 1 *Juan F. Garamendi, 1 Point*

Let

$$\begin{aligned} J: \quad \mathcal{V} &\to \mathbb{R}, \\ u &\mapsto J(u) = \int_\Omega \mathcal{F}\left(\mathbf{x}, u(\mathbf{x}), \nabla u(\mathbf{x})\right) d\mathbf{x} \end{aligned}$$

be a convex energy functional over functions u, where

- $\mathcal{V}$ is a suitable space of functions.
- $\Omega \in \mathbb{R}^d$ is a bounded open domain of the d dimensional euclidean space $\mathbb{R}^d$.
- $u \in \mathcal{V}$, $u : \Omega \to \mathbb{R}$ is a scalar function defined on Ω.
- $\mathbf{x} \in \Omega$ such that $\mathbf{x} = (x_1, .., x_d)$ is the spatial variable and ∇ is the gradient operator such that $\nabla u(\mathbf{x}) = (u_{x_1}, .., u_{x_d})$

(a) (0.25 points) Say in few words which is the fundamental problem in calculus of variations.

(b) (0.25 points) Write the definition of the Gâteaux derivative of $J(u)$ (the directional derivative of J at u in direction of function $h(\mathbf{x})$

(c) (0.25 points) Let

$$\frac{d\,J}{d\,u}$$

be the derivative of J at u. Which is the necessary condition for extremality of J?

(d) (0.25 points) Explain in a few words the relationship between the Gâteaux derivative and the following expressions:

(i)

$$-\text{div}_{\mathbf{x}}\left(\nabla_{\bar{g}}\mathcal{F}\right) + \frac{\partial \mathcal{F}}{\partial u} = 0$$

where

- The divergence $\text{div}_{\mathbf{x}}$ is computed with respect to variable $\mathbf{x}$
- The gradient $\nabla_{\bar{g}}$ is computed with respect to the components of $\bar{g} = \nabla u(\mathbf{x})$

(ii)

$$-\sum_{i=1}^{d} \frac{\partial}{\partial x_i} \frac{\partial \mathcal{F}}{\partial u_{x_i}} + \frac{\partial \mathcal{F}}{\partial u} = 0$$

Problem 2 *Juan F. Garamendi, 1 Points*

For binary segmentation, the model problem can be defined as follows: Let $\omega \subset \Omega$ be an open, positive measured sub-region of the original domain (eventually not connected). If the curve Γ represents the boundary of such a segmentation ω then, in the level set formulation, the (free) moving boundary Γ is the zero level set of a Lipschitz function $\phi : \Omega \to \mathbb{R}$, that is:

$$\Gamma = \{(x, y) \in \Omega : \phi(x, y) = 0\}$$

where

$$\omega = \{(x, y) \in \Omega : \phi(x, y) > 0\}$$

and

$$\Omega \backslash \overline{\omega} = \{(x, y) \in \Omega : \phi(x, y) < 0\}$$

The level set function ϕ can be characterized as a minimum of the following energy functional,

$$J(\bar{c}, \phi) = \int_{\Omega} |DH(\phi)| + \frac{1}{2\lambda} \int_{\Omega} (f - c_1)^2 H(\phi) \ + (f - c_2)^2 (1 - H(\phi)) d\mathbf{x}$$

where $DH(\phi)$ is the distributional gradient of Heaviside function $H(\phi)$, f is a bounded function representing the image (the data), $\bar{c} \in \mathbb{R}^2$, $\bar{c} = (c_1, c_2)$ are a vectorial unknown values representing the classes, and $\lambda \in \mathbb{R}^+$ is a given parameter. The function $H(x)$ represents the Heaviside function, i.e.: $H(x) = 1$ if $x \geq 0$ and $H(x) = 0$ otherwise, and it allows to express the length of Γ by

$$|\Gamma| = \text{Length}(\phi = 0) = \int_{\Omega} |DH(\phi)|$$

where the term $\int_{\Omega} |DH(\phi)|$ denotes, properly, the total variation of the discontinuous function $H(\phi)$ in Ω.

(a) (0.15 points) Why c_1 and c_2 are the mean values of the inside and outside regions?

(b) (0.85 points) Let $c_1, c_2 \in \mathbb{R}$, $c_1 > c_2$, be known values, $\lambda \in \mathbb{R}^+$ be a positive given parameter and $u \in BV(\Omega$ be the (unique) minimum of the Rudin-Osher-Fatemi energy functional

$$J_{rof}(u) = \int_{\Omega} |Du| + \frac{1}{2} \lambda_{rof} \int_{\Omega} |u - f|^2 d\mathbf{x}$$

where $\lambda_{rof} = \frac{c_1 - c_2}{\lambda}$.

Chambolle and Darbon proved[1] that the set $\omega = \{\mathbf{x} \in \Omega | u(\mathbf{x}) > (c_1 + c_2)/2\}$ is a solution of

$$J(\phi) = \int_{\Omega} |DH(\phi)| + \frac{1}{2\lambda} \int_{\Omega} (f - c_1)^2 H(\phi) \ + (f - c_2)^2 (1 - H(\phi)) d\mathbf{x}$$

with $\Gamma = \partial\, \omega = \{\mathbf{x} \in \Omega | \phi(\mathbf{x}) = 0\}$.

Use this information to complete the following algorithm to minimize energy functional $J(\bar{c}, \phi)$ minimizing the Rudin-Osher-Fatemi $J_{rof}(u)$ energy.

[1] A. Chambolle and J. Darbon, *On Total Variation Minimization and surface evolution using parametric maximum flows*. Int. Journal of Computer Vision. 84,3 (2009) pp 288-307

(i) Fix an initial partition ω.

(ii) Minimize the energy $J(\bar{c}, \phi)$ with respect to the constant variables c_1, c_2, i.e. choose c_1 and c_2 as the mean values inside ω and outside ω respectively. Check if $c_1 > c_2$, if it not the case, just exchange names.

(iii)

$\vdots$

(iv)

$\vdots$

Problem 3 *Juan F. Garamendi, 0.5 Points*

Consider the following iterative scheme

- $\bar{x}^k \leftarrow S_i(\bar{x}^{k-1}, \bar{x}^k, \bar{b})$

used to solve the algebraic problem $\mathbf{A}\bar{x} = \bar{b}$, i.e., $\lim_{k\to\infty} \bar{x}^k = \bar{x}$, where $\mathbf{A}$ is a known matrix, $\bar{b}$ is a known vector, $\bar{x}$ is an unknown vector , S_i some function , super-index represents iteration number and $\bar{x}^0$ some initial value. Now consider two versions of S_i: S_1, S_2 with the following behaviour

(a) $e = \bar{x} - S_1(\bar{x}^1, \bar{x}^2, \bar{b})$, $\|e\|_\infty = 10^{-1}$, with e having only high frequencies.

(b) $e = \bar{x} - S_2(\bar{x}^1, \bar{x}^2, \bar{b})$, $\|e\|_\infty = 10^{1}$, with e having only low frequencies.

(a) (0.5 points) Explain in few words which is the best iterative scheme (S_1 or S_2) for embedding it into a multigrid scheme.

Problem 4 Coloma Ballester *0.75 Points*

The following Figure 1 shows 500 measures, $(t_1, y_1), \ldots, (t_{500}, y_{500})$ (where $t_i \in \mathbb{R}$, $y_i \in \mathbb{R}$, $i = 1, \ldots, 500$), which have been taken during the first trimester of 2016, related to the value of stock market shares of a company (a Computer Vision company).

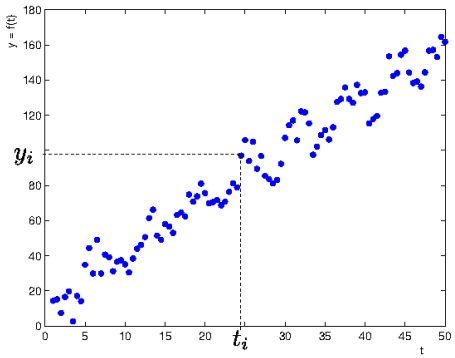

Figure 1: The data.

We assume that there is a functional dependence between t and y, $y = f(t)$, and thus we would like to fit a function f to this given data set. We also assume that f can be modelled as

$$f(t) = at + b + c\cos t,$$

which depends on the (unknown) parameters $a, b, c \in \mathbb{R}$.

(a) Explain the least squares solution to this problem of data fitting, used to determine the parameters $\mathbf{x}^t = [a, b, c]^t$. Write down the expression of the energy $E(\mathbf{x})$ (or $E(a, b, c)$) to be minimized.

(b) Write down the normal equations associated to this problem.

(c) How can you solve the normal equations and determine the parameters $\mathbf{x}^t = [a, b, c]^t$ using the SVD or the pseudoinverse of the matrix associated to your problem?
(Recall that SVD stands for Singular Value Decomposition of a matrix).

Problem 5 Coloma Ballester *1. Points*

Consider the function $f : \mathbb{R}^n \to \mathbb{R}$, $n \in \mathbb{N}$, defined by

$$f(x) = \frac{1}{2}\langle x, Px\rangle - \langle x, q\rangle,$$

for $x \in \mathbb{R}^n$, where P is a $n \times n$ symmetric matrix and $q \in \mathbb{R}^n$. (The notation $\langle \cdot, \cdot \rangle$ stands for the Euclidian scalar product in $\mathbb{R}^n$.)

(a) What condition on the matrix P implies that f is a convex function?

(b) Let $v \in R^n, v \neq 0$. Compute $D_v f(x)$, the directional derivative of f at the point x in the direction v. Use this result to compute $\nabla f(x)$.

(c) Assuming that you are in the situation where f is convex and that P is invertible, verify that the value $x_0 \in \mathbb{R}^n$ where the minimum is attained is $x_0 = P^{-1}q$. Justify the argument you are using.

Problem 6 Coloma Ballester *0.75 Points*

Let A be a $m \times n$ matrix, and $b \in \mathbb{R}^n$, $m, n \in \mathbb{N}$. Consider the problem of minimizing the function $f : \mathbb{R}^n \to \mathbb{R}$ defined by

$$f(x) = \|Ax\|_{\mathbb{R}^m} + \frac{1}{2\lambda}\|x - b\|^2_{\mathbb{R}^n}$$

for $\lambda > 0$. The notation $\|\cdot\|_{\mathbb{R}^k}$ stands for the Euclidean norm in $\mathbb{R}^k$, for $k \in \mathbb{N}$. Note that f is not differentiable when $Ax = 0$.

(a) Write the problem

$$\min_x f(x) \qquad \text{(P)}$$

as a min-max problem. Define also the duality gap.
Hint: Remember the fact that $\|y\|_{\mathbb{R}^k} = \max_{\|\xi\|_{\mathbb{R}^k} \leq 1} \langle y, \xi \rangle_{\mathbb{R}^k}$.

(b) What is the primal-dual problem for problem (P)? Are they equivalent problems?

(c) Define and compute the dual function and the dual problem of problem (P).

Problem 7 Joan Serrat *1 Point*

We saw that binary image denoising could be modeled as a problem of maximum a posteriori inference over the graphical model of Figure 2 below. The goal then was

$$\arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y}) = \arg\max_{\mathbf{x}} p(\mathbf{y}|\mathbf{x})\ p(\mathbf{x})$$

(a) What are $\mathbf{x}$ and $\mathbf{y}$?

(b) What are the names for the $p(\mathbf{y}|\mathbf{x})$ and $p(\mathbf{x})$ terms ?

(c) Which is the "order" of the model ? (can be a number or a word)

(d) The term $p(\mathbf{y})$ does not appear, how comes we can get rid of it ?

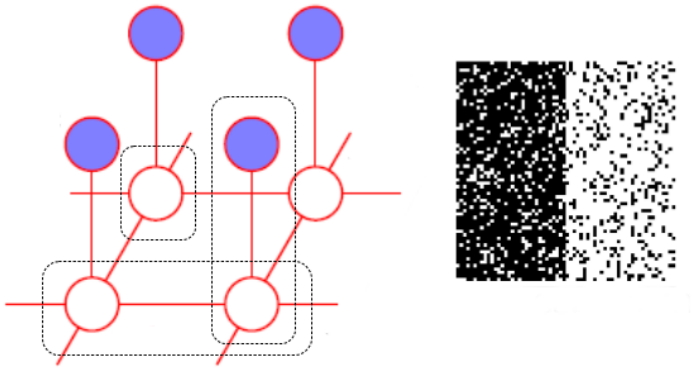

Figure 2: Graphical model for binary image denoising.

Problem 8 Joan Serrat *0.5 Points*

Again with respect to the binary denoising example, we saw that the final equation was

$$\arg\min_{\mathbf{x}} \sum_i \alpha x_i + \sum_{j \in \mathrm{Ne}_i} \beta x_i x_j + \sum_i \gamma x_i y_i$$

where $x_i, y_i \in \{-1, +1\}$ and Ne_i means the neighbors of pixel i. What's false then ? (can be none, one or more choices)

(a) α is related to the mean intensity we expect in a solution

(b) large and positive β makes the solution more smooth

(c) with this formulation we can express our preference for a less smooth solution around image edges

(d) γ is a parameter of the prior

(e) α, β and γ can be learned from training samples

Problem 9 Joan Serrat *0.5 Points*

Which was the solution to the three main difficulties found when trying to optimize the learn the parameters of a graphical model ? Answer writing the pairs of problem-solution labels, like "1-a, 2-b, 3-c".
Difficulties:

(1) $Z(x^i, w)$ or $\mathbb{E}_{y \sim p(y|x^i,w)} \psi(x^i, y)$ impossible to calculate

(2) N small compared to number of parameters, causing overfitting

(3) N large and therefore we have to run belief propagation N times in practice

Solutions:

(a) regularization, assuming w follows a Gaussian distribution

(b) perform stochastic gradient descent

(c) since $\psi(x, y)$ decomposes in factors, we can apply some inference method like belief propagation to compute it

Problem 10 Joan Serrat *0.5 Points*

Consider the graphical model of Figure 3 below where observations are binary images $16 \times 8 = 128$ pixels, that is, $x_i \in \{0,1\}^{128}$, $y_i \in Y = \{\mathrm{a}, \mathrm{b} \ldots \mathrm{z}\}$ (26 lowercase letters), $x = (x_1 \ldots x_9), y = (y_1 \ldots y_9)$.

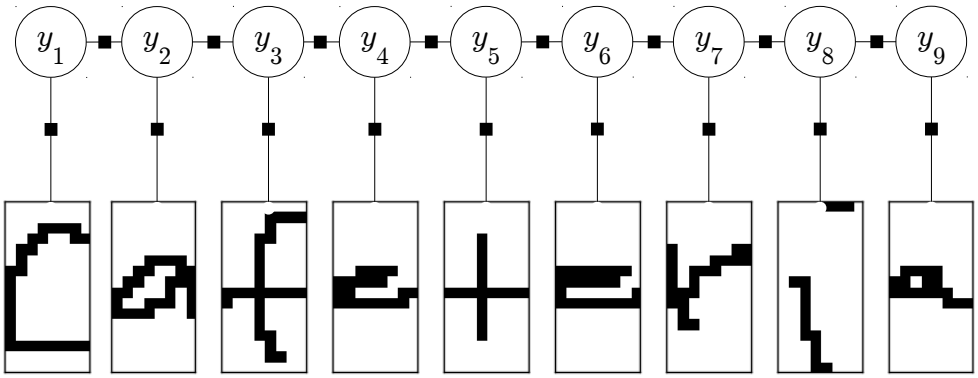

Figure 3.

We want to learn w to later infer a word from a series of binary images of letters as

$$\begin{aligned} y^\star &= \arg\max_{y \in \mathcal{Y}} \langle\, w, \psi(x,y) \,\rangle \\ &= \arg\max_{y \in Y^9} \sum_{i=1}^{9} \sum_{p=\mathrm{a}}^{\mathrm{z}} \sum_{j=1}^{16} \sum_{k=1}^{8} w_{pjk}\, x_{ijk} \;+\; \sum_{i=1}^{8} \sum_{p=\mathrm{a}}^{\mathrm{z}} \sum_{q=\mathrm{a}}^{\mathrm{z}} w_{pq}\, \mathbf{1}_{y_i=\mathrm{p},\, y_{i+1}=\mathrm{q}} \end{aligned}$$

where $\mathbf{1}_{y_i=\mathrm{p},\, y_{i+1}=\mathrm{q}}$ evaluates to 1 if $y_i = \mathrm{p}$ and $y_{i+1} = \mathrm{q}$. In this context,

(a) does it make sense to apply the technique of two stage learning ? why and how ?

(b) what does $w_{p=\mathrm{a},q=\mathrm{b}}$ mean or represent ?

Problem 11 *Oriol Ramos Terrades, 2.5 Points*

The results obtained by the main parties the last general election draw a complex panorama where they will have to negotiate in order to chose the Prime minister of the country. According to public speeches of leaders of parties and their program, experts from the whole country have obtained the following probabilities and conditional probabilities.

First, the A party can agree with both the D party and the I party with the following *a priori* probabilities:

x_1	$p(x_1)$
Agreement with D	0.6
Agreement with I	0.2
None of them	0.2

The D Party doesn't want to reach an agreement with the A party if A agrees with the I party. The conditional probability modeling the vote of the D party is:

x_1	x_2	$p(x_2\|x_1)$
Agreement with D	Vote for A	1
Agreement with D	Vote for B	0
Agreement with D	None	0
Agreement with I	Vote for A	0.2
Agreement with I	Vote for B	0.8
Agreement with I	None	0
None	Vote for A	0.2
None	Vote for B	0.6
None	None	0.4

On the contrary, the C party wishes that the A party will agree with the I party to vote for them. The conditional probability is:

x_1	x_3	$p(x_3\|x_1)$
Agreement with D	Vote for A	0
Agreement with D	None	1
Agreement with I	Vote for A	1
Agreement with I	None	0
None	Vote for A	0.5
None	None	0.5

Moreover, the A party has a lot of problems inside their own organization and depending on the agreement they might reach with they can even vote for the leader of the B party. The conditional probability is:

x_1	x_4	$p(x_4\|x_1)$
Agreement with D	Vote for A	1
Agreement with D	Vote for B	0
Agreement with D	None	0
Agreement with I	Vote for A	0.2
Agreement with I	Vote for B	0.4
Agreement with I	None	0.4
None	Vote for A	0.2
None	Vote for B	0
None	None	0.8

Finally, the probability that they will repeat the elections will basically depend on the votes of the A and the B parties as follows:

x_3	x_4	x_5	$p(x_5\|x_3, x_4)$
Vote for A	Vote for A	0	0.8
Vote for A	Vote for A	1	0.2
Vote for A	Vote for B	0	1
Vote for A	Vote for B	1	0
Vote for A	None	0	0
Vote for A	None	1	1
None	Vote for A	0	0
None	Vote for A	1	1
None	Vote for B	0	1
None	Vote for B	1	0
None	None	0	0
None	None	1	1

where $x_5 = 0$ means that parties want avoid to repeat the elections.

Assume that we have observed $x_5 = 0$ and $x_2 =$"Vote for B". We wonder if we can infer the exact probability $p(x_4|x_2, x_5)$.

a) A Bayesian net can model this problem. Draw it and write the domain of each random variable (0.5 points).

b) Write the joint distribution given by the factorization defined by the Bayesian net (0.5 points).

c) We define the factor function, $\phi_a(x_3, x_4, x_5) = p(x_5|x_3, x_4)$, to model the interaction between these three variables. This factor defines a 3-order clique. Write the matrices $\phi_a(x_3, \cdot, x_5)$ for each possible value of x_4 (0.5 points).

d) Define the remaining factor functions and draw the corresponding factor graph (0.5 points)?

e) Does Belief Propagation (BP) provide exact estimates of marginals? compute the message $m_{3\leftarrow d}(x_3)$ if it provides exact inference. Otherwise explain why you can not apply BP (0.5 points).