Module: M2. Optimization and inference techniques for Computer Vision **Final exam**
Date: December 4th, 2018
Teachers: Juan F. 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, 2 Points

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) (0.5 points) Say in a few words which is the fundamental problem in calculus of variations.
The fundamental problem of the calculus of variations is to find the extremum (maximum or minimum) of the functional $J(u)$ with respect to u.

(b) (1 points) Let $J(u)$ be (strictly) convex. Explain in few words the difference between minimizing $J(u)$ using 1) calculus of variations and 2) the backpropagation strategy
For finding the minimum using calculus of variations we can procedure in two ways:

(i) Finding the necessary condition that has to be satisfied by the minimum of $J(u)$, that is

$$\frac{d\,J}{d\,u} = 0$$

i.e., compute the Euler-Lagrange equation

(ii) Apply a gradient descent scheme to iteratively find the minimum

Figura 1: `Matrix` and `tensor` words are both alias for A matrix. `rhs` word is an alias for b vector.

In both cases, the derivative of J w.r.t. u has to be computed and discretized. This ends up in a (maybe linear) algebraic system of equations that has to be solved.

On the other hand, using backpropagation, the energy functional $J(u)$ has to be decomposed in atomic operations with known derivatives and the chain rule can be applied to compute the gradient within a gradient descent scheme.

(c) (0.5 points) What if $J(u)$ is not convex ?

If $J(u)$ is not convex, it will have several minimums, i.e. the Euler-Lagrange equation will have multiple solutions or no solution, and the solution reached by any gradient descent scheme (using back propagation or directly calculus of variations) will depend on the starting point.

Problem 2

Juan F. Garamendi, $min(1, 2.a + 2.b)$ Point

You have decided to use the TensorFlow (open source) library for minimizing $J(u)$ using calculus of variations. At the end, you have an algebraic system of equations $A\bar{x} = b$ where A is a well-conditioned mxm square matrix and $\bar{x}$ and $\bar{b}$ are vectors of size m and $\bar{x}$ is the unknown. In figure 1 you can find screenshots from tensorflow documentation.

(a) (1 points) Which function from figure 1 you should try first? Explain in a few words why. Because A is well conditioned and square, we should try first solving directly the linear system of equations with tf.linalg.solve.

(b) (1 points) Bonus track for curious students: Explain the difference between solving the inpainting problem as you solve in the project, and the backprogpagation tensorflow implementation given in class. What are the advantages and disadvantages of each one?

In the project we solved the Euler-Lagrange equation for the inpainting problem, i.e., we computed the derivative of the functional with respect to u and we discretized the resulting Euler-Lagrange equation, that this ended up in an algebraic system of equations that we solved. On the contrary, with the back propagation, the problem was solved decomposing the problem into atomic operations with known derivatives. These derivatives (and the chain rule) were used to compute the derivative of the functional on each step of a gradient descent scheme. The advantage of the first method is speed. The code is much more faster, but the disadvantage is that you have to know how to compute the derivative of the energy functional with respect the unknown and how to numerically solve the resulting Euler-Lagrange equation. On the other hand, using backpropagation inside a descent gradient scheme only requieres to know the derivative of the atomic operations but it is slow to converge, as usual in gradient descent schemes.

Problem 3 Coloma Ballester *0.75 Points*

Consider the constrained minimization problem

$$\begin{aligned} \min_{x_1,x_2} \quad & -x_1 + 3x_2 \\ \text{subject to} \quad & \tfrac{1}{2}x_1 - x_2 + 2 \geq 0 \\ & x_1 - 3 \leq 0, \\ & x_1 + x_2 + 3 \geq 0. \end{aligned}$$

(a) Sketch the set of constraints of the problem. Is it a convex set? (0.2 *points*)

This is a problem of the form $\min_{\mathbf{x}\in C} f(\mathbf{x})$, where C is the convex set given by the closed triangle of the following figure:

(b) Write the KKT optimality conditions for the problem. (0.4 *points*)

The Lagrange dual function associated to the problem is

$$\mathcal{L}(x_1, x_2, \lambda_1, \lambda_2) = -x_1 + 3x_2 - \lambda_1(\frac{1}{2}x_1 - x_2 + 2) - \lambda_2(3 - x_1) - \lambda_3(x_1 + x_2 + 3)$$

Thus, the KKT optimality conditions are

$$\begin{cases} \nabla_x \mathcal{L}(x, \lambda_1, \lambda_2) = \begin{bmatrix} -1 \\ 3 \end{bmatrix} + \lambda_1 \begin{bmatrix} -\frac{1}{2} \\ 1 \end{bmatrix} + \lambda_2 \begin{bmatrix} 1 \\ 0 \end{bmatrix} - \lambda_3 \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \\ \frac{1}{2}x_1 - x_2 + 2 \geq 0 \\ 3 - x_1 \geq 0 \\ x_1 + x_2 + 3 \geq 0 \\ \lambda_1 \geq 0, \lambda_2 \geq 0, \lambda_3 \geq 0 \\ \lambda_1(\frac{1}{2}x_1 - x_2 + 2) = 0 \\ \lambda_2(3 - x_1) = 0 \\ \lambda_3(x_1 + x_2 + 3) = 0. \end{cases}$$

(c) Check if any of the points $(x_1, x_2) = (0, -3)$ and $(x_1, x_2) = (3, -6)$ could be the solution of the problem using the KKT conditions. (0.15 *points*)

Both points belong to the feasible set C but only the second one, $(3, -6)$, satisfies the KKT optimality conditions. It is the solution of the problem as, in this case, they also are sufficient conditions.

Problem 4 Coloma Ballester *0.75 Points*

Let us consider the vectors $\mathbf{x}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$, the $n \times n$ real matrix A, $\alpha > 0$, and the following minimization problem:

$$\min_{\mathbf{x}} \|A\mathbf{x}\| + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle$$

(a) Is this a convex function? Why? (0.15 *points*)

Yes, the objective function, $\|A\mathbf{x}\| + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle$, is convex as it consists of the sum of convex functions: the composition of a norm and a linear function (both convex functions), plus a quadratic function, plus a linear function.

(b) Write an equivalent min-max problem and the resulting iterations of a primal-dual algorithm to solve it. (0.4 *points*)

Using that $\|Ax\| = \max_{\xi \in C} \langle Ax, \xi \rangle$, where $C = \{\xi \in \mathbb{R}^n : \|\xi\| \leq 1\}$, we have that:

$$\|A\mathbf{x}\| + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle = \max_{\xi \in C} \left(\langle Ax, \xi \rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle \right),$$

Then:

$$\min_x \left(\|A\mathbf{x}\| + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle \right) = \min_x \max_{\xi \in C} \left(\langle Ax, \xi \rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle \right).$$

The function

$$\mathcal{L}(x, \xi) = \langle Ax, \xi \rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle,$$

depending on the primal variables x and on the dual variables ξ, is convex with respect x (for each $\xi \in C$ fixed) and concave with respect to ξ (for each $x \in \mathbb{R}^n$ fixed). Therefore, the duality gap is zero. Thus, there exists a saddle point and the original Primal problem, the Primal-Dual problem, and the Dual problem are three equivalent problems.

The Primal-Dual problem is solved by alternating a a projected gradient ascent step for the variable ξ, and gradient descent step for the variable x:

$$\xi^{k+1} = P_C(\xi^k + \tau \nabla_\xi \mathcal{L}(x^k, \xi^k))$$

$$x^{k+1} = x^k - \theta \nabla_x \mathcal{L}(x^k, \xi^{k+1}).$$

where $P_C(v) = \frac{v}{\max\{1, \|v\|\}}$ is a projector over C (for any vector vector $v \in \mathbb{R}^n$).
In our case, the 'partial gradients' of $\mathcal{L}$, $\nabla_x \mathcal{L}$ and $\nabla_\xi \mathcal{L}$ with respect to x and ξ, respectively, are given by

$$\nabla_x \mathcal{L}(x, \xi) = A^t \xi + \alpha(x - b) + c$$

$$\nabla_\xi \mathcal{L}(x, \xi) = Ax$$

Finally, we find a solution by iterating the following update equations

$$\xi^{k+1} = P_C(\xi^k - \tau A x^k)$$

$$x^{k+1} = x^k - \theta \left(A^t \xi^{k+1} + \alpha(x^k - b) + c \right),$$

(c) Outline the dual algorithm to solve this problem? (0.2 *points*)

The dual function is $g_D(\xi) = \mathcal{L}(x_0(\xi), \xi)$, where

$$x_0(\xi) = \arg\min_x \mathcal{L}(x, \xi) = \arg\min_x \left(\langle Ax, \xi \rangle + \frac{\alpha}{2}\|\mathbf{x} - \mathbf{b}\|^2 + \langle \mathbf{x}, \mathbf{c} \rangle \right).$$

The minimizer $x_0(\xi)$ is the solution of $\nabla_x \mathcal{L}(x, \xi) = 0$, which is

$$x_0(\xi) =$$

Substituting $x_0(\xi)$ one obtains the dual function:

$$g_D(\xi) = \mathcal{L}(x_0(\xi), \xi) = .$$

Finally the dual problem is

$$\max_{\xi \in C} g_D(\xi) = \max_{\xi \in C} .$$

(which is a quadratic problem with constraints, where we have eliminated the primal variable, and therefore could be solved with a projected gradient ascent).

Problem 5 J.Serrat *0.5 Points*

We saw that binary image denoising could be modeled as a problem of maximum a posteriori inference over the graphical model of figure 2. 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 ?

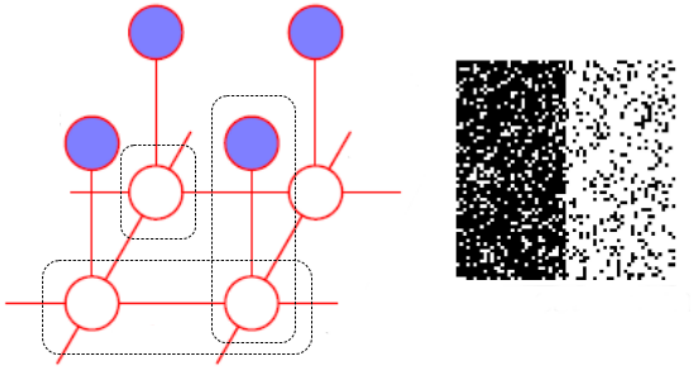


Figura 2: graphical model for binary image denoising

(a) $\mathbf{x}$ is the sought clean image, $\mathbf{y}$ is the observation or noisy image we have

(b) $p(\mathbf{y}|\mathbf{x})$ is the likelihood, $p(\mathbf{x})$ the prior, in the Bayesian speak

(c) order two or pairwise

(d) $p(\mathbf{y})$ is the evidence, it appears when applying the Bayes theorem to reverse probabilities (from $p(\mathbf{x}|\mathbf{y})$ to $p(\mathbf{y}|\mathbf{x})$) and we can discard it because we are maximizing over $\mathbf{x}$ only

Problem 6 J.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

b and c and d

Problem 7 J.Serrat *0.5 Points*

Consider the graphical model of figure 3 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), $x = (x_1 \ldots x_9), y = (y_1 \ldots y_9)$.

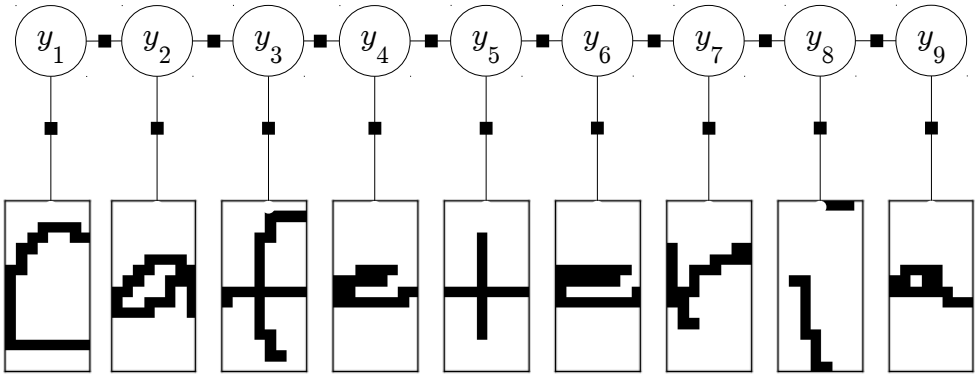


Figura 3

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 ?

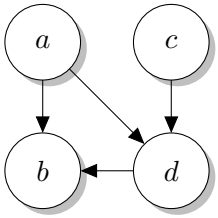
(b) what does $w_{p=\mathrm{a},q=\mathrm{b}}$ mean or represent ?

(a) Yes, it makes sense to apply two-stage learning. There are about 3000 unary parameters and only 26*26 pairwise, and it is possible that the learning could be driven by the first group. Instead of one weight per pixel and possible label, we could train some classifier (like an SVM) on the binary images and use the vector of 26 probabilities as features. Then the number of unary parameters would be the same as pairwise.

(b) compatibility of 'b' right after 'a', high if this is a frequent pair in our training set

Problem 8 *Oriol Ramos Terrades, 0.5 Points*

Given the following Bayesian network (BN):



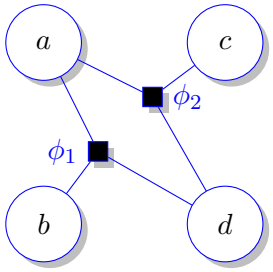
a) Write the joint distribution given by the conditional probabilities inferred from the BN.

Solution:

$$p(a,b,c,d) = p(b|a,d)p(d|a,c)p(a)p(c)$$

b) Draw a factor graph derived from the BN. Also, write the definition of its factor functions in terms of the conditional probabilities used in the joint distribution.

Solution:



$$\phi_1(a,b,d) = p(b|a,d)p(a)$$
$$\phi_2(a,c,d) = p(d|a,c)p(c)$$

Problem 9 *Oriol Ramos Terrades, 1 Point*

Say whether the next statements are true (**T**) or false (**F**) [Correct: +0.25, Incorrect: -0.25, unanswered: 0 points].

a) Belief propagation infers exact marginals in acyclic directed graphs.

b) The complexity of the message passing algorithm on a chain model is $O(N^2K)$, where N is the number of identically distributed random variables, X_i, and K is the number of states of X_i.

c) Samples that are consecutively generated by the Gibbs sampling method are independently distributed.

d) The Normalized importance sampling method generates unbiased estimators.

Solution:

a) True.

b) False.

c) False.

d) False.