Module: M2. Optimization and inference techniques for Computer Vision **Final exam**
Date: November 30th, 2017
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, 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 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) (0.25 points) Write the definition of the Gâteaux derivative of $J(u)$ (the directional derivative of J at u in the direction of a function $h(\mathbf{x})$).

$$\left.\frac{d\,J}{d\,u}\right|_h = \lim_{\alpha \to 0} \frac{J(u + \alpha h) - J(u)}{\alpha}$$

(c) (0.5 points) Say in a few words what are the Euler-Lagrange Equations for a given convex functional.
In the calculus of variations, the Euler–Lagrange equation, is a second-order partial differential equation whose solutions are the functions for which a given functional is stationary, and, for a convex functional, when it has a minimum.

Problem 2 *Juan F. Garamendi, 1 Point*

The OpenCV library (Open source Computer Vision Library) has a class called SVD. In the cv::SVD class reference documentation is said the following:

```
Class for computing Singular Value Decomposition of a floating-point matrix.
The singular Value Decomposition is used to blah, blah, blah, blah
```

(a) (0.5 points) Say at least two problems in which SVD is useful.

There are many problems for which SVD is usefull: Least-square problem, to solve under(over)-determined linear systems, invert matrices (using pseudo-inverse)....

(b) (0.5 points) Say in a few words how to solve the problem (or at least compute an approximated solution to the problem) $A\bar{x} = b$ where A is a mxn matrix and $\bar{x}$ and $\bar{b}$ are vectors of size n and m, respectively, with $m > n$, and $\bar{x}$ is the unknown.

This algebraic system of equations is overdetermined, so, probably it may not be possible to satisfy all restrictions, but we can compute an approximated solution in the sense of least squares problem, that states

$$\min_{\bar{x}} \|A\bar{x} - \bar{b}\|^2$$

The equation whose solutions are the minimum of the previous minimization problem are the normal equations

$$\bar{x} = (A^t A)^{-1} A^t \bar{b}$$

That can be solved using the SVD to compute $(A^t A)^{-1}$

Problem 3 *Juan F. Garamendi, 1. 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) $e = \bar{x} - S_1(\bar{x}^1, \bar{x}^2, \bar{b})$, $\|e\|_\infty = 10^{-10}$, with e having only high frequencies.

(b) $e = \bar{x} - S_2(\bar{x}^1, \bar{x}^2, \bar{b})$, $\|e\|_\infty = 1000$, 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.

In a multigrid scheme the iterative scheme S, used to obtain the approximated solution $\bar{x}^k$, must have a remarkable smoothing effect on the error $e = \bar{x} - \bar{x}^k$. This is because the residual equation $\mathbf{A}\bar{e} = \bar{r}$, being $\bar{r} = \bar{b} - \mathbf{A}\bar{x}^k$ the residual, will be solved in a coarser level than the original algebraic problem $\mathbf{A}\bar{x} = \bar{b}$, so it is important that $\bar{e}$ in the coarser level represents well the error $\bar{e}$ at the finer level, and this is only possible if $\bar{e}$ does not have high frequencies (We want to avoid the aliasing effect), i.e. the best scheme is S_2.

(b) (0.5 points) Explain the relationship between Gauss-Seidel and Multigrid and in which moment of a multigrid alogrithm is used Gauss-Seidel. Is Gauss-Seidel the best scheme to use with multigrid?

Gauss-Seidel is used in a multigrid scheme and in some problems as Poisson problem as error smoother. Just with a few iterations of the G-S we get an approximated solution. This approximated solution can be very far from the solution, but the high frequencies of the error has been removed. G-S is not the only smoother scheme, the goodness and effectiveness depend on the problem, so, the smoother scheme has to be chosen for every problem and sometimes G-S is not the best option.

(a) (i) What is a convex set? (ii) Give an example of a convex set. (0.2 points)

Solution: (i) A set $C \in \mathbb{R}^N$ (or $C \in V$, where V is a suitable space of functions) is convex if given $x, y \in C$, the segment joining x and y is contained in C. That is, given any $x, y \in C$, then

$$[x, y] := \{tx + (1-t)y : t \in [0,1]\} \subset C.$$

(ii) Half-spaces ($\{x \in \mathbb{R}^N : \langle a, x\rangle + b \geq \lambda\}$ where $a \in \mathbb{R}^N$, $b \in \mathbb{R}$, $\lambda \in \mathbb{R}$) are convex sets.

(b) (i) What is a convex function (or a convex functional)? (ii) Give an example of a convex function $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ (that is, $f(\mathbf{x}) \in \mathbb{R}$ for any vector $\mathbf{x} \in \mathbb{R}^N$), and an example of a convex functional $J : V \longrightarrow \mathbb{R}$ (that is, $J(u) \in \mathbb{R}$ for any function $u \in V$, where V is a suitable space of functions). Give also an example of a non-convex functional (or function, if you prefer). (0.4 points)

(i) A function or functional $f : C \to \mathbb{R}$ is convex if

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),$$

for any $x, y \in C$ and any $t \in [0,1]$ (notice that C needs to be a convex set in order to always have $tx + (1-t)y \in C$).

(ii) A norm $f(x) = \|x\|$ in $\mathbb{R}^N$ is a convex function.

Also: Let $b \in \mathbb{R}^N, c \in \mathbb{R}$. Then, $f(x) = \langle x, b\rangle + c$ is a convex function in $\mathbb{R}^N$.

Also: Let A be an $n \times n$ matrix, $b \in \mathbb{R}^N, c \in \mathbb{R}$. $f(x) = \langle Ax, x\rangle + \langle x, b\rangle + c$ is convex iff A is positive definite.

An example of convex functional is

$$J(u) = \frac{1}{2\lambda}\int_\Omega |u - f|^2 dx + \frac{1}{2}\int_\Omega |\nabla u|^p dx$$

where f is a known noisy image, u is an unknown image, $p = 2$ or $= 1$, and λ is a given parameter that controls the trade-off between the data fidelity term and the smoothness or regularity term. $J(u)$ is the sum of two which are the composition of convex functions (which is convex) and norm-type functions, and all the norms are convex functions.

An example of a non-convex functional is

$$E(u) = \frac{1}{2\lambda}\int_\Omega |u - f|^2 dx - \frac{1}{2}\int_\Omega |\nabla u|^p dx$$

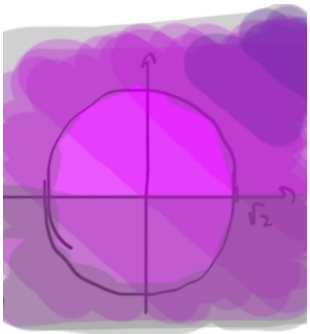
with $p = 2$ or $= 1$.

(c) Consider the following constrained optimization problem (0.4 points)

$$(\mathcal{P}) \begin{cases} \min\limits_{x_1, x_2} f(x_1, x_2) = x_1 + x_2 \\ \text{subject to} \\ \quad 2 - x_1^2 - x_2^2 \geq 0 \\ \quad x_2 \geq 0 \end{cases}$$

(i) This is a problem of the form $\min\limits_{\mathbf{x}\in C} f(\mathbf{x})$. Draw a picture of the constraint set C.

This is a problem of the form $\min\limits_{\mathbf{x}\in C} f(\mathbf{x})$, where C is the convex set given by the upper semi-disc of the following figure:



(ii) What are the associated Karush-Kuhn-Tucker (KKT) optimality conditions?
The Lagrange dual function associated to the problem is

$$\mathcal{L}(x_1, x_2, \lambda_1, \lambda_2) = x_1 + x_2 - \lambda_1(2 - x_1^2 - x_2^2) - \lambda_2 x_2$$

Thus, the KKT optimality conditions are

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

Problem 5 Coloma Ballester *1 Points*

Choose and answer only one of the following two options:

Option 5.A. 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 $\|y\|_{\mathbb{R}^k} = \max_{\|\xi\|_{\mathbb{R}^k} \leq 1} \langle y, \xi \rangle_{\mathbb{R}^k}$.
Using that $\|Ax\|_{\mathbb{R}^m} = \max_{\xi \in C} \langle Ax, \xi \rangle_{\mathbb{R}^m}$, where $C = \{\xi \in \mathbb{R}^m : \|\xi\|_{\mathbb{R}^m} \leq 1\}$, we have that:

$$f(x) = \max_{\xi \in C} \left(\langle Ax, \xi \rangle_{\mathbb{R}^m} + \frac{1}{2\lambda}\|x - b\|^2_{\mathbb{R}^n} \right),$$

Then:

$$\min_x f(x) = \min_x \max_{\xi \in C} \left(\langle Ax, \xi \rangle_{\mathbb{R}^m} + \frac{1}{2\lambda}\|x - b\|^2_{\mathbb{R}^n} \right),$$

The duality gap is the difference

$$DG = \min_{x \in R^n} \max_{\xi \in C} \left(\langle Ax, \xi \rangle + \frac{1}{2\lambda}\|x - b\| \right) - \max_{\xi \in C} \min_{x \in R^n} \left(\langle Ax, \xi \rangle + \frac{1}{2\lambda}\|x - b\| \right).$$

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

$$\mathcal{L}(x, \xi) = \langle Ax, \xi \rangle + \frac{1}{2\lambda}\|x - b\|^2,$$

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), then $DG = 0$

and the three problems (the Primal problem (P), the Dual problem, and the Primal-Dual problem) are equivalent.
The Primal-Dual problem is

$$\min_{x\in R^n}\max_{\xi\in C}\left(\langle Ax,\xi\rangle+\frac{1}{2\lambda}\|x-b\|\right)=\max_{\xi\in C}\min_{x\in R^n}\left(\langle Ax,\xi\rangle+\frac{1}{2\lambda}\|x-b\|\right)$$

(c) Define and compute the dual function and the dual problem of problem (P).
The dual function is $g_D(\xi)=\mathcal{L}(x_0(\xi),\xi)$, where

$$x_0(\xi)=\arg\min_x\mathcal{L}(x,\xi)=\arg\min_x\left(\langle Ax,\xi\rangle+\frac{1}{2\lambda}\|x-b\|\right).$$

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

$$x_0(\xi)=b-\lambda A^t\xi.$$

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

$$g_D(\xi)=\mathcal{L}(x_0(\xi),\xi)=\langle Ab,\xi\rangle_{\mathbb{R}^m}-\frac{\lambda}{2}\|A^t\xi\|^2_{\mathbb{R}^m}.$$

Finally the dual problem is

$$\max_{\xi\in C}g_D(\xi)=\max_{\xi\in C}\langle Ab,\xi\rangle_{\mathbb{R}^m}-\frac{\lambda}{2}\|A^t\xi\|^2_{\mathbb{R}^m}.$$

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

Option 5.B. Let A be an $m\times n$ matrix, $b\in\mathbb{R}^m$. Consider the problem

$$\begin{aligned}&\min\|x\|^2 &&\text{(P)}\\ &\text{subject to } Ax=b.\end{aligned}$$

(a) Write problem (P) as a min-max problem and define the duality gap.
$Ax=b$ gives m equality constraints on x: $(Ax)_i=b_i$, $i=1,\cdots,m$. Therefore, we introduce m Lagrange multipliers (or dual variables), $\nu_1,\ldots,\nu_m$, and we construct the Lagrangian function $\mathcal{L}(x,\nu)=f(x)-\sum_{i=1}^m\nu_i((Ax)_i-b_i)=\langle x,x\rangle-\langle\nu,Ax-b\rangle=\langle x,x\rangle-\langle A^t\nu,x\rangle+\langle\nu,b\rangle$, where $\nu=(\nu_1,\ldots,\nu_m)^t\in R^m$. Therefore

$$\min_{\text{subject to } Ax=b}\langle x,x\rangle=\min_{x\in R^n}\max_{\nu\in R^m}\mathcal{L}(x,\nu)$$

The duality gap is the difference

$$DG=\min_{x\in R^n}\max_{\nu\in R^m}\mathcal{L}(x,\nu)-\max_{\nu\in R^m}\min_{x\in R^n}\mathcal{L}(x,\nu).$$

(b) Define and compute the dual function of problem (P).
In our case, $DG=0$ because $\mathcal{L}(x,\nu)$ is convex on x (f is convex and $d_i(x)=(Ax)_i-b_i$ are linear constraints) and $\mathcal{L}(x,\nu)$ is concave on ν. Therefore we can change min-max by max-min:

$$\min_{\text{subject to } Ax=b}\langle x,x\rangle=\min_{x\in R^n}\max_{\nu\in R^m}\mathcal{L}(x,\nu)=\max_{\nu\in R^m}\min_{x\in R^n}\mathcal{L}(x,\nu)=\max_{\nu\in R^m}g_D(\nu),$$

where $g_D(\nu)$ ($= \mathcal{L}(x_0(\xi), \xi)$ and $x_0(\xi) = \arg\min_{x \in R^n} \mathcal{L}(x, \nu)$) is the dual function. We compute the dual function from $\nabla_x \mathcal{L}(x, \nu) = 0$ (indeed, $\mathcal{L}(x, \nu)$ is a quadratic function of x, and for each ν there is a unique minimizer $x_0(\nu)$. The minimizer is the solution of $\nabla_x \mathcal{L}(x, \nu) = 0$). In our case, $2x - A^t\nu = 0$, which gives $x_0(\nu) = \frac{1}{2}A^t\nu$. Then,

$$g_D(\nu) = \mathcal{L}(x_0(\nu), \nu) = -\frac{1}{4}\langle A^t\nu, A^t\nu\rangle + \langle \nu, b\rangle$$

(c) Write down the dual problem.
$\max_{\nu \in R^m}(-\frac{1}{4}\langle A^t\nu, A^t\nu\rangle + \langle \nu, b\rangle)$.

Problem 6

Joan Serrat, 0.5 Points

The three main problems found when trying to learn the parameters of a graphical model in the probabilistic formulation are :
Problems:

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

(2) N large and therefore we have to run belief propagation N times

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

We saw that they could be overcomed by :

(a) regularization, assuming w follows a Gaussian distribution

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

(c) perform stochastic gradient descent

Now, which problem is solved by what ?

a) 1-a, 2-b, 3-c

b) 1-b, 2-c, 3-a

c) 1-c, 2-a, 3-b

d) 1-b, 2-a, 3-c

Correct answer : b

Problem 7

J.Serrat *0.5 Points*

Consider the graphical model of Figure 1 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), $x = (x_1 \ldots x_9), y = (y_1 \ldots y_9)$.

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

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

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, what's **false** ? Can be one or more.

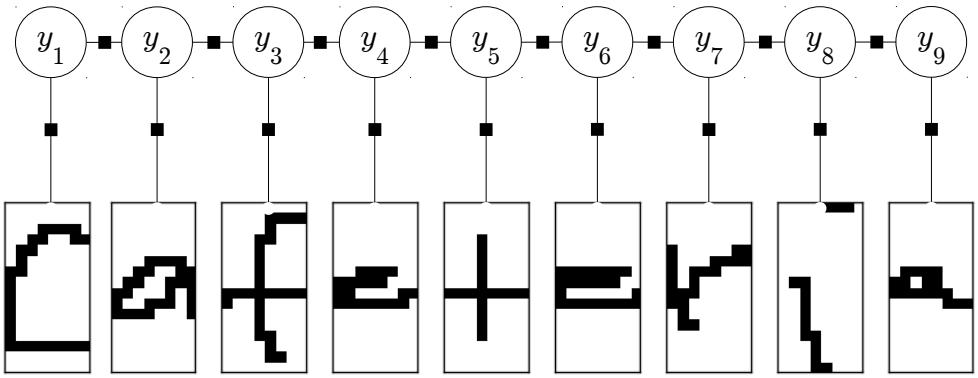


Figura 1

a) the total number of unary parameters to learn is 26^2

b) there are $16 \times 8 \times 26$ pairwise parameters

c) $w_{p=\mathrm{a},q=\mathrm{b}}$ means the compatibility of labeling p as letter a and q as letter b, being p and q any pair of nodes in the chain

d) w_{pq} are the parameters of the prior term

Correct answers : a, b and c

Problem 8 *Joan Serrat, 0.5 Points*

The two-stage training is a technique (mark the true one)

a) to learn such that then makes inference faster

b) to learn a balanced set of unary and pairwise coefficients

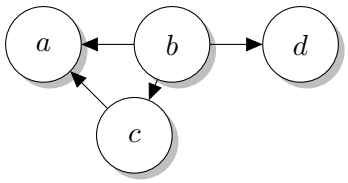
c) that implies not learning at all the unary coefficients

d) to speed up learning only

Correct answer : b

Problem 9 *Oriol Ramos Terrades, 0.5 Points*

Given the following Bayesian network:



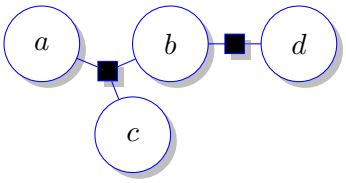
a) Write the joint distribution according to the conditional probabilities inferred from the Bayesian network.

Solution:

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

b) Draw a factor graph derived from it.

Solution:



Problem 10 *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 infer exact marginals in loopy graphical models.

b) The complexity of loopy belief propagation depends on the order of the highest clique.

c) Samples generated by the Metropolis-Hasting algorithm are always accepted.

d) Sampling methods provides exact inference on tree-based graphical models.

Solution:

a) False

b) True

c) False

d) False