ECSE626 Statistical Methods in Computer Vision
Assignment #1

Andrei Chubarau, 260581375, andrei.chubarau@mail.mcgill.ca

March 3, 2018

1 Part I: Regularization

Firstly, we define the overdetermined system of equations

$$Ax = b \tag{1}$$

The solution to the above system is given in Equation 2.

$$x^* = \min_x ||Ax - b||^2 + \alpha||x||^2 \tag{2}$$

a) The parameter α controls the enforced regularization, or the smoothness of the solution, by influencing the amount of penalty induced from model complexity, i.e. the magnitude of the solution vector x.

b) Smaller values for the parameter α are preferred, because this minimizes the amount of regularization, thus giving the "best" solution with less bias. While regularizing the problem is a priority, it is undesirable to introduce too much bias and smoothness.

c) Minimization of the defined system can be done using a gradient method. This is done by following the negative of the gradient until the gradient becomes zero, which corresponds to reaching the local minimum (not necessarily global). Thus, we want to follow the negative of the gradient of the functional $M^\alpha(x, b)$. The gradient of the system defined in Equation 2 is given by the following:

$$\nabla_X M^\alpha(x, b) = 2A^T(Ax - b) + 2\alpha^2 X \tag{3}$$

The solution to this then can be directly computed as

$$x = (A^T A + \alpha^2 I)^{-1} A^T b \tag{4}$$

The above can also be solved iteratively using a gradient iteration approach, for instance with Newton's method of the form:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \tag{5}$$

d) Given our knowledge about the mathematical properties of this problem and the stated assumptions about the system, we know that the L2 error is directly affected by the choice of the regularization constant α. At the same time, α influences the smoothness of the solution, by giving proportionate preference to solution vectors with smaller magnitude. Thus, we can define an optimal α value by specifying a balance between solution smoothness and the maximum L2 error bound.

For the cases where α approaches zero, from the definition of the L2 error, we see that the upper bound on the error goes to infinity:

$$E_{L2}\{||x^* - X_0^*||\}^{\frac{1}{2}} \leq \sigma/2\alpha \tag{6}$$

$$\lim_{\alpha \to 0} \sigma/2\alpha = \infty \tag{7}$$

Therefore, when $\alpha \to 0$, the upper bound on the L2 error does not exist, which will lead to unstable solutions - the problem returns to being ill-posed. Since the system is unstable, the effect of the presence of noise will have a strong influence on the solution; this is what we want to avoid by employing regularization.

An optimization search can be run to find the optimal value of α. This can be done by applying cross-validation, to evaluate different values of α and find a convenient trade-off; otherwise, a gradient approach can be taken to find the optimal hyperparameter setting.

e) Bayesian approach solves the problem by focusing on maximizing a probability distribution given the original system. To be more specific, Bayesian approach would use the following formulation:

$$p(x|b) = \frac{p(b|x)p(x)}{p(b)} \tag{8}$$

The solution to the above is given by maximizing $p(x, b)$; this corresponds to the Maximum A Posteriori (MAP) solution which minimizes the probability of error. Since this requires a complete model of the system (involving distributions for $p(x)$, $p(b)$, and $p(b|x)$), this might be difficult to solve, and so the formulation of the Bayesian approach is further explored in the next part where a convenient optimization technique is presented.

f) The equivalence of the two approaches implies that they produce the same solution, which we will investigate. In the case of Tikhonov regularization, we have the system

$$p(x|b) = \frac{\exp(-\beta M^{\alpha}(x, b))}{X} \tag{9}$$

We claim that Bayesian and Tikhonov approaches are equivalent if the conditional probability $p(x|b)$ has a Gibbs distribution, defined in Equation 10.

$$p(x|b) = \frac{\exp(-\beta E(x, b))}{X} \tag{10}$$

Thus, we associate the system defined in Equation 9 with the energy functional E(x,b) of the Gibbs distribution, which when expanded results in the following:

$$p(x|b) = \frac{\exp(-\beta p_U(Ax, b)) \exp(-\beta\alpha\Omega[x]}{X} \tag{11}$$

We demonstrate equivalence by decomposing Equation 11 into the following components modeled as probabilities in the Bayesian approach:

$$p(x|b) = \frac{\exp(-\beta p_U(Ax, b))}{X_1} \tag{12}$$

$$p(x|b) = \frac{\exp(-\beta\alpha\Omega[x]}{X_2} \tag{13}$$

$$p(x|b) = \frac{X}{X_1 X_2} \tag{14}$$

2 Part II: Information Theory

2.1 Entropy

The entropy of a distribution of a random variable X is defined as:

$$H(X) = -\sum_{x \epsilon X} p(x) log(x) \tag{15}$$

To compute the entropy of a digital grayscale image, I first compute the histogram of the color values c of the input image; this is done by finding unique color values and counting them in the image. The probability for each possible color c_i is then computed as the total count of pixels with color c_i (similar to a histogram) divided by the total number of pixels. Finally, entropy is calculated according to Equation 15.

2.1.1 Solutions

a) The entropy of the image 000.png is 3.2213.

b) The entropy of a random noise array with noise amplitude equal 20 is 5.2719.

c) The entropy of the image injected with noise is 6.3088. Note that injecting noise into images potentially results in values that are outside of the range [0-255]; no clamping was applied in order to maintain the uniform nature of noise.

d) Uniform random noise with amplitude in the range [0-200] is injected into the input image as described in c). Figure 1 shows the relationship between noise amplitude and entropy of the resulting combined image. The observed trend suggests that the entropy increases as more noise is injected into an image. This is as expected since random noise should have high entropy; mixing noise into an image would increase the entropy of said image proportionally to the randomness and amplitude of the noise. Moreover, the combined images have higher entropy than that of the individual inputs; this can be explained by the randomness of noise in addition to the existing variation of the input image.

2.2 Mutual Information & KL Divergence

Mutual information (MI), joint entropy (JE), and Kullback-Leibler (KL) divergence can be computed as per Equations 16, 17, and 18, respectively.

$$I(X;Y) = \sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \tag{16}$$

$$H(X,Y) = -\sum_{x \epsilon X} \sum_{y \epsilon Y} p(x,y) log(p(x,y)) \tag{17}$$

$$D(p||q) = \sum_{x \epsilon X} p(x) log \frac{p(x)}{q(x)} \tag{18}$$

My implementation for computing joint probability of the involved distributions is based on finding unique values in the image arrays. I do not use any built-in histogram functions, so

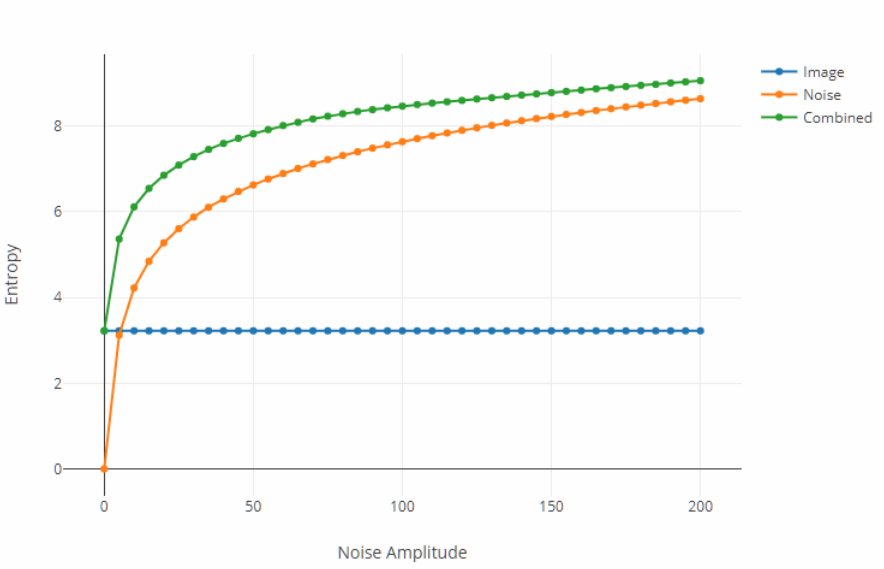


Figure 1: Entropy of an image influenced by injection of random noise

that I have full freedom in this step. Firstly, we merge two images to create what essentially is a third image with two color channels, one for each input image. This array contains pairs of pixels; each unique pair $c_1 2$ is then identified and its corresponding probability is computed as total number of occurrences of that pair divided by total number of occurrences.

MI, JE, and KL divergence values are then computed over the defined range given the distributions, in such a way that combinations with zero probability are never included in the calculation (since they are irrelevant and have no effect).

In our case, joint entropy describes the uncertainty in the joint distributions of colors $c \in C$ given two images. Mutual information is a somewhat similar metric, in the sense that they are inversely proportional.

2.2.1 Solutions

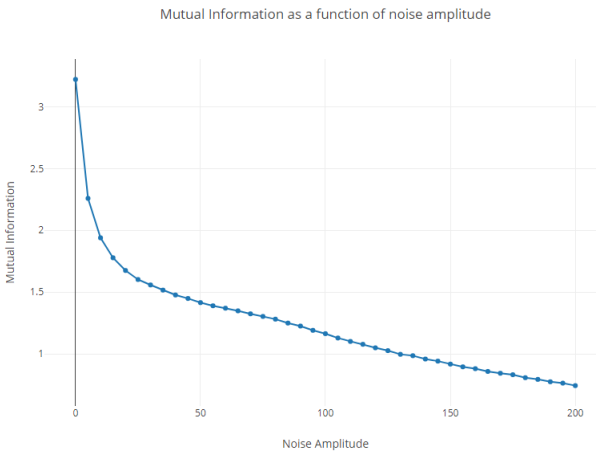


Figure 2: Mutual Information as a function of injected noise amplitude

a) Mutual information of image 000.png with a random noise array is 0.0983, which is essentially almost zero, as expected.

b) Individual entropies of image 000.png and a random noise array are 3.2213 and 5.2712, respectively; joint entropy is 8.3942. This results is consistent with the values of individual entropy and mutual information; it is equal to the sum of the two entropies minus mutual information.

c) Kullback-Leibler divergence values are as follows: i) between two instances of noise: 5.3603e-4, ii) between image 000.png and a noise array: 1.7650, iii) between image and a noisy image 2.0428; mutual information value between the image and the noisy image is 1.6738. Kulback-Leibler divergence is a measure of how one probability distribution diverges from another. Since two instances of noise values have similar distributions, it is expected that their KL divergence is small; for similar reasons, KL between the noise array and the image should be high. Lastly, KL divergence between an image and a noisy image are higher than either of the two previous results, which is also as expected; while the original distribution

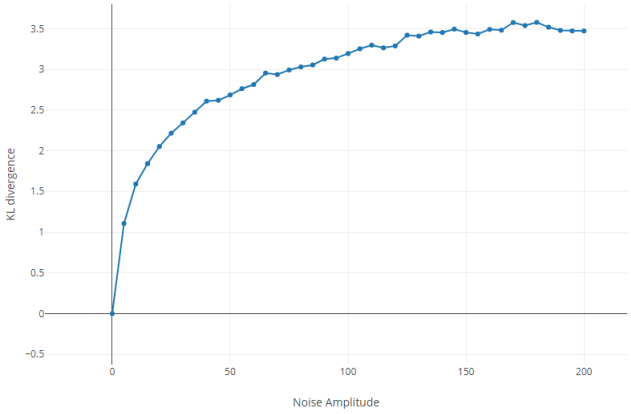


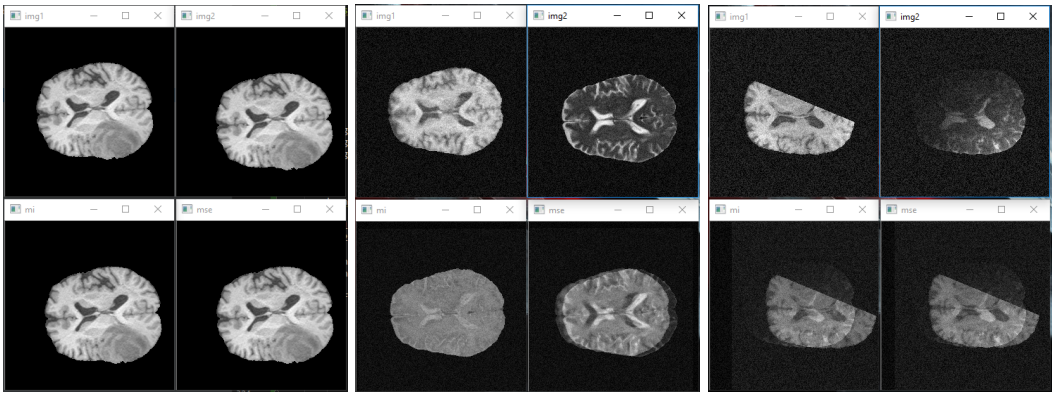
Figure 3: Kullback-Leibler divergence as a function of injected noise amplitude

d) Figure 2 shows the trend of Mutual Information between an image and a noisy version of same image as a function of noise amplitude. The trend indicates that MI is decreasing as more noise is injected, which is explained by the fact that the noise degrades the image, reducing its original content, thereby reducing the amount of entropy that can be correlated between the two versions.

Note that pixel values were not clamped to range [0-255] after noise was injected.

e) Figure 3 shows the trend of Kullback-Leibler divergence between an image and a noisy version of same image as a function of noise amplitude. The trend indicates that the KL divergence increases for the tested range and plateaus towards higher amplitudes of noise. This is as expected, because the mathematical definition of KL divergence implies higher and higher divergence for distributions that are more and more different, as is the case for an image affected by more and more noise. Nevertheless, as higher amplitude noise is injected, we see that the divergence between the two distributions stabilizes; this is simply because at that point we are comparing the original image to a more and more noisy version.

I also investigated KL divergence in the reversed order (comparing noisy image to image as opposed to comparing image to noisy image). Since KL divergence is not symmetric, this does not give exactly the same result, although a similar trend is still observable.



(a) 000.png and 001.png (b) 002.png and 003.png (c) 004.png and 005.png

Figure 4: Optimal translations for three pairs of images given by two different metrics: i) MI (bottom left), ii) MSE (bottom right)

2.3 Simple Image Registration

a) The optimal translation for both metrics is (15, 12), see Figure 4a.

b) The optimal translation given by MSE is (9, -8), while by MI is (10, 3), see Figure 4b. The two metrics do not give the same optimal result. The presence of noise in the images makes it so that MI is a more reliable metric. MSE is too sensitive to noisy inputs, and thus cannot evaluate optimal translations as effectively.

c) The optimal translation given by MSE is (0, 19), while by MI is (1, 31), see Figure 4c. The two input images are quite different from each other. Firstly, the colors appear inverted; secondly, there is observable occlusion; lastly, the images are degraded by severe noise. Neither of the metrics result in optimal translations. There are significant errors in both cases since the input imagery is difficult for both of the used optimization techniques. MSE suffers from presence of noise, while MI is not as effective due to occlusion.