We consider the problem of image denoising; that is, given an image that has been distorted by noise in the image formation process, we would like to estimate the true image (without noise). One can imagine several applications for this problem:

- turbulence imagery (discuss in class)
- ultrasound images (discuss in class)
- etc...

0.1 A Note About Models

In order to be able to denoise an image, we need to know how the image is formed and how the noise is generated. Typically, this would be application dependent since the formation of images is different and the noise generated would be different in various imaging scenarios. For example, the noise generated due to atmospheric conditions in satellite imagery is different from the noise generated from imaging an organ using an ultrasound probe. In this course, we will create models for no specific application (simply because of time), but try to create generic models usually for natural images. Note this will naturally be less powerful than having application specific models. Nevertheless, we will learn to construct and reason about the models, which will enable you to construct models for specific applications when additional information is known.

0.2 Additive Noise Model

For image denoising, we start with a simple model that is, as we shall see, the basis for many popular denoising algorithms. We will denote the image to denoise as, $I$, and we define it as a function $I : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$, where $\Omega$ is some (compact) subset of the plane (usually a rectangle), and for now we assume the image is grayscale. We say $I$ is the observation.

**Definition 1** The additive noise image formation model is

$$I = u + \eta,$$

(1)

where $I : \Omega \to \mathbb{R}$ is the observed image, $u$ is the (true image or denoised image), and $\eta$ is the noise. The true image, $u$, belongs to some specified class of functions denoted $\mathcal{U}$, and $\eta$ is the noise, which is a random variable.
Remark 1  
• Models as the one above, that allow you to generate a sample image, are called generative models. We have to choose the class of functions $U$ and choose the distribution of the noise, $\eta$, and we will do so in due course.

• We sometimes refer to $u$ as the state (aka hidden variable or latent variable), and $I$ as the measurement, in terminology from the controls literature.

0.3 Bayesian Approach

We are now going to use a methodology, called the Bayesian approach, whereby we can estimate the probability of a true image $u \in U$ given that we have observed the image $I$. Once this is done, we can choose the $u$ that maximizes this probability to be our denoised image. Let us denote by $P$ the probability of an event.

We are going to estimate $P(u|I)$ using Bayes’ Theorem:

$$P(u|I) = \frac{P(I|u)P(u)}{P(I)} = \frac{P(I|u)P(u)}{\sum_u P(I|u)P(u)}. \quad (2)$$

Remark 2  
• $P(u|I)$ is called the posterior

• $P(I|u)$ is called the likelihood, a measure of how likely the image is given knowledge of the state

• $P(u)$ is called the prior, which encodes our knowledge about the states possible

For more details and philosophy of the Bayesian approach, see [1].

0.4 Computing the Posterior Using Additive Noise Model

0.4.1 Likelihood

Using our additive noise model, and specification of the noise, we are going to compute $P(I|u)$. We are going to make the assumption that the noise, $\eta(x)$, is a Gaussian random variable with mean 0 and variance $\sigma^2$ for each $x \in \Omega$; we write $\eta(x) \sim \mathcal{N}(0, \sigma)$. Further we will assume that $\eta(x)$ and $\eta(y)$ are independent when $x \neq y$. In otherwords, $\{\eta(x)\}_{x \in \Omega}$ are iid (independent, identically distributed). Formally, we write

$$p(\eta(x) = t) \propto \exp \left( -\frac{t^2}{2\sigma^2} \right). \quad (3)$$

where $p$ denotes a probability density.

Remark 3 The Gaussian assumption ...

Given our additive noise model, we can now compute $p(I(x)|u(x))$:

$$p(I(x) = t|u(x)) = p(u(x) + \eta(x) = t|u(x)) = p(\eta(x) = t - u(x)) \propto \exp \left( -\frac{(t - u(x))^2}{2\sigma^2} \right); \quad (4)$$

we simply write for convenience,

$$p(I(x)|u(x)) \propto \exp \left( -\frac{(I(x) - u(x))^2}{2\sigma^2} \right). \quad (5)$$
By independence of the noise process, we can now calculate $p(I|u)$:

$$p(I(x_1), \ldots, I(x_n)|u(x_1), \ldots, u(x_n)) = p(\eta(x_1) = I(x_1) - u(x_1), \ldots, \eta(x_n) = I(x_n) - u(x_n))$$

$$= \prod_{i=1}^{n} p(\eta(x_i) = I(x_i) - u(x_i)) \propto \exp\left(-\frac{\sum_{i=1}^{N}(I(x_i) - u(x_i))^2}{2\sigma^2}\right)$$

(6)

where $x_1, \ldots, x_n \in \Omega$ are distinct. By using a limiting argument from calculus, we can write

$$p(I|u) \propto \exp\left(-\frac{1}{2\sigma^2} \int_{\Omega} (I(x) - u(x))^2 \, dx\right)$$

(7)

where the integration above is with respect to the area element in $\Omega$.

### 0.4.2 Prior

Above we have calculated the likelihood, and in order to compute the posterior, we need to specify the prior, $p(u)$, and also the class $\mathcal{U}$. Note that the class $\mathcal{U}$ is a class of functions that are to represent the class of true or denoised images. We are left with deciding the definition of a denoised image. This is a hard question, and as mentioned before, it depends on the specific type of images we are considering, and the particular application.

We are going to assume that a denoised image is smooth compared to images that are noisy. One measure of smoothness is derived by noting that smooth functions have derivatives that have low values. Indeed for many natural imagery, it is true that image derivatives are near zero in a large portion of the image (e.g. look around the class room, it is composed of many regions where the intensity is nearly constant). Now this is not true everywhere (for example, between two different objects there is an edge, and that has large derivatives); however, edges cover a small area region of the image compared to nearly constant patches. Therefore, we assume that $u_x(x), u_y(x) \sim \mathcal{N}(0, \sigma_p)$ where $u_x$ and $u_y$ denote partial derivatives of $u$ in the $x$ and $y$ directions. We will in addition assume the independence of $u_x(z_1), u_y(z_2)$ when $z_1 \neq z_2$. I have no defense of this last assumption (in fact it is often false, indeed, if $z_1$ and $z_2$ are close, then $u_x(z_1)$ and $u_y(z_2)$ are usually correlated) other than saying that it will lead to tractable computational algorithms. Therefore, completing analysis similar to the previous section, we find

$$p(u) \propto \exp\left(-\frac{1}{2\sigma^2} \int_{\Omega} |\nabla u(x)|^2 \, dx\right)$$

(8)

where $\nabla u(z) = (u_x(z), u_y(z))$ is the gradient of $u$, and $| \cdot |$ denotes the standard Euclidean norm. Let’s look at what the prior probability states. It says that the more unsmooth the function (large values of $\nabla u$, the less likely the function is to be a denoised image. We choose the class of functions $\mathcal{U}$ to be square integrable functions:

$$\mathcal{U} = \left\{ u : \Omega \to \mathbb{R} : \int_{\Omega} |\nabla u(x)|^2 \, dx < \infty \right\},$$

(9)

which is often denoted as $H^1(\Omega; \mathbb{R})$.

---

1. Actually, the additional condition $\int_{\Omega} |u(x)|^2 \, dx < +\infty$ is needed to call the space $H^1$. The space $H^1$ is called a Sobolev space (and the 1 indicates that the first derivative is square integrable). Sobolev spaces play a fundamental role in the theory of linear PDE.
0.4.3 Maximum A-Posteriori (MAP) Estimate

Now that our likelihood and prior probabilities are known, we are going to estimate the denoised (true) image given that we have the observed (or noisy image). One way of doing this, is known as maximum a-posteriori (MAP) estimation method. That is, we find the denoised image \( \hat{u} \) among those in \( U \) that maximizes the posterior distribution. That is,

\[
\hat{u} = \arg \max_u p(u|I),
\]

which by using Bayes’ Theorem is equivalent to

\[
\hat{u} = \arg \max_u p(I|u)p(u),
\]

(notice the denominator in Bayes’ Theorem is ignored since it does not depend on \( u \), which has been integrated out). We have computed the two probabilities in the previous sections. Before we seek the maximum, we note that since \( \log \) is an increasing function, we can equivalently solve the maximization problem

\[
\hat{u} = \arg \max_u \log (p(I|u)p(u)) = \arg \max_u \log p(I|u) + \log p(u);
\]

further, maximizing an argument is equivalent to minimizing negative of that argument, and so

\[
\hat{u} = \arg \min_u -\log p(I|u) - \log p(u).
\]

We refer to the right hand argument as the energy; we define

\[
E(u) = -\log p(I|u) - \log p(u).
\]

Thus, the MAP estimate is the \( u \in U \) that minimizes the above energy. Using the expressions for the likelihood and priors above, we find that

\[
E(u) = \frac{1}{2\sigma^2_l} \int_\Omega (I(x) - u(x))^2 \, dx + \frac{1}{2\sigma^2_p} \int_\Omega |\nabla u(x)|^2 \, dx.
\]

For simplicity, we choose \( \sigma_l \) and \( \sigma_p \) so that

\[
E(u) = \int_\Omega (I(x) - u(x))^2 \, dx + \alpha \int_\Omega |\nabla u(x)|^2 \, dx
\]

where \( \alpha > 0 \). Often the first term is called the data fidelity term, and the second is called the regularization term. In the next lecture, we are going to construct algorithms to find the optimizer of \( u \).

References