ECE 592 – Topics in Data Science

Test 5: Sparse Signal Processing – Fall 2022

December 9, 2022

Question 1 (Homeworks and final projects.)
Among the homeworks or final projects you peer-graded (not your final project), describe something that made an impression on you. Why was it interesting? What did you learn?

Solution. Any reasonable response would be fine here.

Question 2 (Sparse recovery with sparsifying basis.)
You are designing a compressive signal acquisition system. The system will acquire unknown inputs, $x \in \mathbb{R}^N$, using linear measurements, $y = Ax + z$, where $A \in \mathbb{R}^{M \times N}$ is a measurement matrix, $z \in \mathbb{R}^M$ is Gaussian noise, and you will recover x from y, A, and possible statistical properties of x.

(a) Suppose that you have plenty of training data, meaning many instances of x. Could you use the training data to identify a good sparsifying transform W for x? How?

Solution. Use PCA to find a sparsifying transform, W, that optimizes the amount of energy we can pack into a small number of dimensions.

(b) We now have a sparsifying transform, meaning that $\theta = Wx$ will tend to be sparse. (In the context of this part, you can state any assumptions about θ that you want.) How would you reconstruct x from y, A, W, and possible statistical knowledge (your assumptions) about x? Note that we are asking you how to reconstruct x, not θ.

Solution. Any reasonable assumptions for θ would be reasonable, for example a sparse signal where the nonzero coefficients are Gaussians. We now have $y = Ax + z = AW\theta + z$. Reconstruct or estimate θ by $\widehat{\theta}$ using any reasonable sparse recovery method, for example LASSO or AMP. Next, $\widehat{x}$ is obtained by multiplying $\widehat{\theta}$ by W, i.e., $\widehat{x} = W\widehat{\theta}$.

Question 3 (Approximate message passing with ℓ_1 penalty.)
Consider a sparse recovery problem, $y = Ax + z$, where $x \in \mathbb{R}^N$ is an unknown input vector, $A \in \mathbb{R}^{M\times N}$ is a measurement matrix, $z \in \mathbb{R}^M$ is Gaussian noise, and the goal is to recover x from y, A, and possible statistical properties of x.
Many sparse recovery approaches try to minimize the mean squared error (MSE) between the unknown input, x, and our estimate, $\widehat{x}$. What if we want to minimize the ℓ_1 error,

$$\widehat{x} = \arg\min_{x'} E[\|x' - x\|_1 | y, A]? \tag{1}$$

(Recall that $\|\cdot\|_1$ denotes the ℓ_1 norm of a vector, which is the sum of absolute values of individual vector entries.) In Test 4, we saw that linear regression with an ℓ_1 penalty could be performed by perturbing an existing solution for conventional regression with an ℓ_2 penalty. In this question, we will try to solve (1) using approximate message passing (AMP).
The key idea underlying AMP is that the linear inverse problem, $y = Ax + z$, is decoupled into a scalar denoising problem,

$$v = x + w, \tag{2}$$

where $v \in \mathbb{R}^N$ are noisy measurements, and $w \in \mathbb{R}^N$ is additive white Gaussian noise with zero mean and variance σ^2. AMP performs several denoising steps to estimate x from v, where the variance σ^2 tends to diminish until some noise floor is approached. A conditional expectation denoiser is often used,

$$\widehat{x} = E[X|V = v], \tag{3}$$

where $\widehat{x}$ is our estimate for the unknown input, x, because conditional expectation minimizes the mean squared error and thus drives down noise in the scalar denoising problem. However, our goal is to find the estimate $\widehat{x}$ that minimizes the expected ℓ_1 error (1), not the squared error, hence a conditional expectation denoiser in the last iteration seems inappropriate.
To perform ℓ_1 regression, we propose a variant of AMP, which is comprised of 2 parts. In Part 1, we reduce the noise level in our scalar denoising problem (2) for several iterations. To do so, a conditional expectation denoiser (3) reduces the noise level (σ^2) in each iteration quickly. Next, Part 2 performs ℓ_1 denoising,

$$\widehat{x} = \arg\min_{x'} E[\|x' - x\|_1 | v]. \tag{4}$$

Below, you will derive both denoisers for an example probability density function (pdf).
To keep the derivations simple, we will consider a scalar denoising problem where X is a scalar random variable whose unknown value is x, and we estimate x from a noisy measurement, $v = x + w$. The noise is a Gaussian random variable, $W \sim \mathcal{N}(0, \sigma^2)$ whose pdf obeys $f(W = w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{w^2}{2\sigma^2}}$. Again, to keep the derivations simple, the random variable X will take values -1 or $+1$, each with probability 0.5.

(a) For Part 1, compute the conditional expectation denoiser (3) for X. (Derive a scalar denoiser.)

Solution. We compute the conditional expectation, (3),

$$\begin{aligned}
E[X|V = v] &= 1 \cdot \Pr(X = 1|V = v) + (-1) \cdot \Pr(X = -1|V = v) \\
&= \frac{\Pr(X = 1) f(V = v|X = 1)}{f(V = v)} - \frac{\Pr(X = -1) f(V = v|X = -1)}{f(V = v)} \\
&= \frac{\Pr(X = 1) f(V = v|X = 1) - \Pr(X = -1) f(V = v|X = -1)}{\Pr(X = 1) f(V = v|X = 1) + \Pr(X = -1) f(V = v|X = -1)}.
\end{aligned}$$

Because $\Pr(X=-1)=\Pr(X=+1)$, these cancel out in the numerator and denominator,

$$E[X|V=v]=\frac{f(V=v|X=1)-f(V=v|X=-1)}{f(V=v|X=1)+f(V=v|X=-1)}.$$

Next,

$$f(V=v|X=-1)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(v+1)^2}{2\sigma^2}}, \quad f(V=v|X=+1)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(v-1)^2}{2\sigma^2}}.$$

Therefore,

$$\begin{aligned} E[X|V=v] &= \frac{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(v+1)^2}{2\sigma^2}}-\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(v-1)^2}{2\sigma^2}}}{\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(v+1)^2}{2\sigma^2}}+\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(v-1)^2}{2\sigma^2}}} \\ &= \frac{e^{-\frac{(v+1)^2}{2\sigma^2}}-e^{-\frac{(v-1)^2}{2\sigma^2}}}{e^{-\frac{(v+1)^2}{2\sigma^2}}+e^{-\frac{(v-1)^2}{2\sigma^2}}}. \end{aligned}$$

(b) For Part 2, we want an ℓ_1 denoiser, which in this scalar case minimizes the absolute value of the error,

$$\widehat{x}=\arg\min_{x'} E[|x'-x||v]. \tag{5}$$

We will show in steps that $\widehat{x}$, the optimal ℓ_1 estimate for x, is the median of the posterior, $f(x|v)$, meaning that

$$\Pr(x\geq\widehat{x}|v), \Pr(x\leq\widehat{x}|v)\geq 0.5.$$

Let us discuss this median estimator. First, the probability for x to be smaller or larger than $\widehat{x}$ depends on v (2), hence we condition on v in both probability expressions. Second, when X is a continuous valued random variable, $\Pr(X=\widehat{x}|v)=0$, hence

$$\Pr(x\geq\widehat{x}|v)=\Pr(x\leq\widehat{x}|v)=0.5. \tag{6}$$

If $f(x)$ contains probability masses (deltas in the pdf), then we could have $\Pr(X=\widehat{x}|v)>0$, in which case (6) might not be correct. To simplify our question, we will assume that X is continuous valued, hence you will show the simplified result, (6).
To show this result (6), define the function,

$$\mu(x',v)=E[|x'-x||V=v],$$

which is the expected ℓ_1 (in our scalar variable case, absolute value) error between x and $\widehat{x}$, conditioned on the scalar channel v. In the optimal ℓ_1 denoising function, (5), $\widehat{x}$ minimizes $\mu(x',v)$. We now analyze $\mu(x',v)$,

$$\begin{aligned} \mu(x',v) &= \int_{t=-\infty}^{+\infty}|x'-t|f_X(X=t|V=v)dt \\ &= \int_{t=-\infty}^{x'}|x'-t|f_X(X=t|V=v)dt+\int_{t=x'}^{+\infty}|x'-t|f_X(X=t|V=v)dt. \end{aligned}$$

Explain why we partitioned the first integral where t was within the range $(-\infty, +\infty)$ into integrals with ranges $(-\infty, x')$ and $(x', +\infty)$. Specifically, how does $|x' - t| f_X(X = t|V = v)$ change between these ranges?

Solution. In the range $(-\infty, x')$, $|x' - t| = x' - t$. In the range $(x', +\infty)$, $|x' - t| = -(x' - t)$. Therefore, we partition the integral in a way that the absolute value term can be expressed in 2 different ways.

(c) Again, $\widehat{x}$ minimizes $\mu(x', v)$. Because X is continuous valued, we can compute the derivative of $\mu(x', v)$ with respect to x',

$$
\begin{aligned}
\frac{\partial}{\partial x'}\mu(x', v) &= \frac{\partial}{\partial x'} \int_{t=-\infty}^{x'} |x' - t| f_X(X = t|V = v)dt + \frac{\partial}{\partial x'} \int_{t=x'}^{+\infty} |x' - t| f_X(X = t|V = v)dt \\
&= \int_{t=-\infty}^{x'} [+1] f_X(X = t|V = v)dt + \int_{t=x'}^{+\infty} [-1] f_X(X = t|V = v)dt \\
&= \int_{t=-\infty}^{x'} f_X(X = t|V = v)dt - \int_{t=x'}^{+\infty} f_X(X = t|V = v)dt,
\end{aligned}
$$

where $[-1]$ and $[+1]$ in red font are the derivatives of $|x' - t|$ in the 2 integrals. (The derivative of an integral may include terms caused by the bounds of the interval changing, but this is not a concern here.) At the minimum, this derivative is zero,

$$
\int_{t=-\infty}^{x'} f_X(X = t|V = v)dt = \int_{t=x'}^{+\infty} f_X(X = t|V = v)dt. \tag{7}
$$

How does (7) relate to what we wanted to show, (6)?

Solution. The integral on the left hand side equals $\Pr(X \leq x'|V = v)$. The integral on the right hand side equals $\Pr(X \leq x'|V = v)$. Moreover, the sum of the 2 probabilities is 1, because $\Pr(X = x'|V = v) = 0$ (recall that X is assumed to be a continuous valued random variable). Combining these observations, we obtain (6).

(d) Recall that X takes values -1 or $+1$, each with probability 0.5. What is the optimal ℓ_1 denoising function, $\eta(v)$? (Hint: your actual answer will have a very simple form.)

Solution. In this part, X is not continuous valued, and we must use the more complicated expression,

$$
\Pr(x \geq \widehat{x}|v), \Pr(x \leq \widehat{x}|v) \geq 0.5.
$$

Because X can only take the values -1 or $+1$, the posterior estimator is either -1 or $+1$. Owing to symmetry, it is -1 when $v < 0$, else it is $+1$. (We ignore the edge case when $v = 0$.)