A Derivation of the Soft-Thresholding Function

Ivan Selesnick
Polytechnic Institute of New York University

April 28, 2009. Last edit: September 28, 2018

These notes show the derivation of non-linear soft-thresholding function for signal denoising. The soft-thresholding function can be used for denoising by applying it to the transform-domain representation, provided the transform yields a sparse representation of the signal. For example, wavelet transforms provide sparse representations of piece-wise smooth signals, and the short-time Fourier transform (STFT) provides sparse representations of oscillatory signals (like speech).

1 Derivation of the Soft-Threshold Function

We assume that a signal of interest has been corrupted by additive noise, i.e.

$$g = x + n \tag{1}$$

where n is white zero-mean Gaussian noise independent of the signal x. We observe g (a noisy signal), and wish to estimate the noise-free signal x as accurately as possible. In the transform domain (eg, wavelet, STFT, etc), the problem can be formulated as

$$y = w + n \tag{2}$$

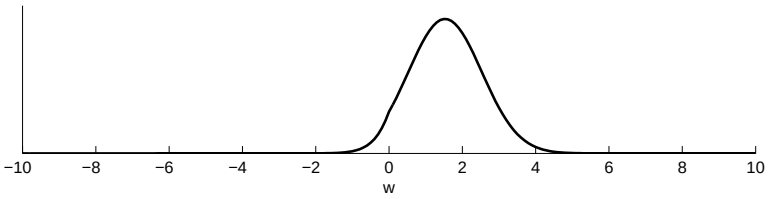
where y is the noisy coefficient (eg, wavelet coefficient), w is the noise-free coefficient and n is noise, which is again zero-mean Gaussian (any linear transform of a zero-mean Gaussian random signal results in a zero-mean Gaussian random signal). If the transform is orthogonal, then the noise in the transform domain has the same correlation function as the original noise in the signal domain; therefore, when the transform is orthogonal, white noise in the signal domain becomes white noise in the transform domain.

Our goal is to estimate w from the noisy observation y. The estimate will be denoted as $\hat{w}$. Because the estimate depends on the observed (noisy) value y, we also denote the estimate as $\hat{w}(y)$. We will use the maximum a posteriori (MAP) estimator. The MAP estimator is based on the probability density function (pdf) of w. Specifically, given an observed value y, the MAP estimator asks what value of w is most likely? That is, the MAP estimator looks for the value of w where the probability of w is highest; it looks for the peak value. Therefore, the MAP estimator is defined as

$$\hat{w}(y) = \arg\max_{w} p_{w|y}(w|y) \tag{3}$$

where 'arg max' is the value of the argument where the function has its maximum. The pdf $p_{w|y}(w|y)$ is the distribution of w given a specific value y.

Conditional pdf, $p_{w|y}(w|y)$, for some value of y.



The MAP estimate $\hat{w}$ is the point where the pdf has its peak.

To find the value of w where $p_{w|y}(w|y)$ has its peak, note that

$$p_{w|y}(w|y) = \frac{p_{w,y}(w,y)}{p_y(y)}$$

and

$$p_{y|w}(y|w) = \frac{p_{w,y}(w,y)}{p_w(w)}$$

so rearranging terms we get

$$p_{w|y}(w|y) = \frac{p_{y|w}(y|w)\, p_w(w)}{p_y(y)}.$$

(This is Bayes rule.) Therefore, one gets

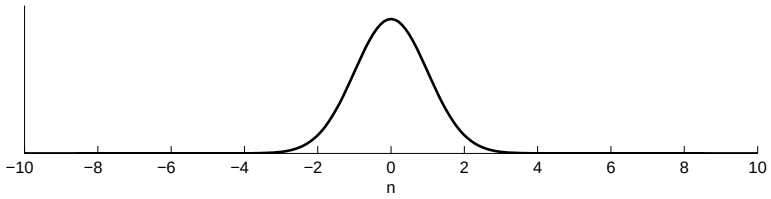
$$\hat{w}(y) = \arg\max_{w} \frac{p_{y|w}(y|w)\, p_w(w)}{p_y(y)}.$$

Because the term $p_y(y)$ does not depend on w, the value of w that maximizes right-hand side is not influenced by the denominator. Therefore the MAP estimate of w is given by

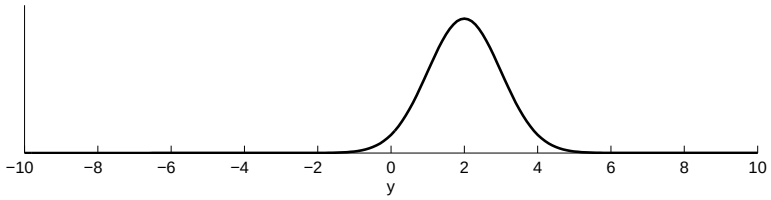
$$\hat{w}(y) = \arg\max_{w} \left[p_{y|w}(y|w) \cdot p_w(w) \right].$$

The conditional pdf $p_{y|w}(y|w)$ can be found by noting that given w, we have that $y = w + n$ is the sum of w and a zero-mean Gaussian random variable. For example, if n is a zero-mean Gaussian random variable, then $2 + n$ is a Gaussian random variable with mean 2 and the pdf will be centered around 2. Similarly, if w is known, then $w + n$ is a Gaussian random variable with mean w and the pdf will be centered around w. Therefore, $y = w + n$ is Gaussian with mean w. That is: $p_{y|w}(y|w) = p_n(y - w)$.

The pdf, $p_n(n)$, of a zero-mean Gaussian random variable.



The pdf, $p_n(y-2)$, of a Gaussian random variable with mean 2.



Therefore, $p_{y|w}(y|w) = p_n(y-w)$ and the estimate can be written as:

$$\hat{w}(y) = \arg\max_w \left[p_n(y-w) \cdot p_w(w)\right]. \tag{4}$$

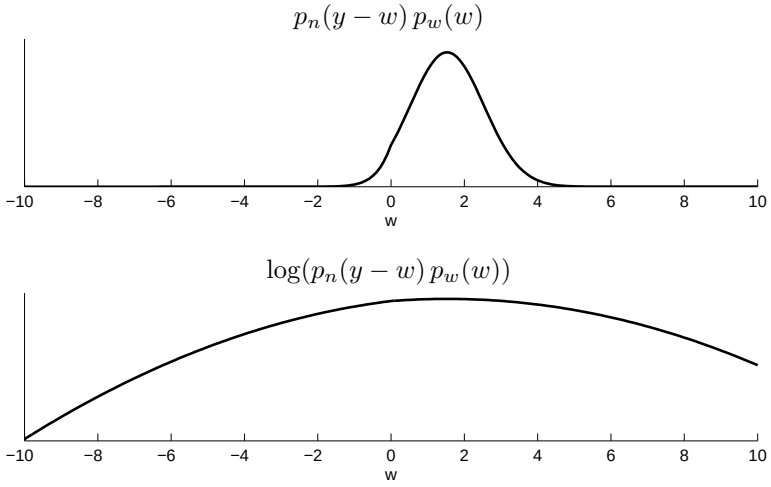
Note that the value w where a function $F(w)$ has its maximum is not changed when a monotonic function $G()$ is applied to the function. In other words, the value of w maximizing $F(w)$ is also the value of w maximizing $G(F(w))$. The logarithm is monotonic, so (4) is also equivalent to

$$\hat{w}(y) = \arg\max_w \left[\log(p_n(y-w)\, p_w(w))\right] \tag{5}$$

or equivalently,

$$\hat{w}(y) = \arg\max_w \left[\log(p_n(y-w)) + \log(p_w(w))\right]. \tag{6}$$

Using the logarithm just simplifies the subsequent differentiation step.



The location of the peak is unchanged by taking the logarithm of the function.

We have assumed the noise is zero mean Gaussian with variance σ_n,

$$p_n(n) = \frac{1}{\sigma_n\sqrt{2\pi}} \cdot \exp\left(-\frac{n^2}{2\,\sigma_n^2}\right). \tag{7}$$

By using (7), (6) becomes

$$\hat{w}(y) = \arg\max_w \left[-\frac{(y-w)^2}{2\,\sigma_n^2} + \log(p_w(w))\right]. \tag{8}$$

Let's define $f(w) = \log(p_w(w))$. Then we get

$$\hat{w}(y) = \arg\max_w \left[-\frac{(y-w)^2}{2\,\sigma_n^2} + f(w)\right]. \tag{9}$$

We can therefore obtain the MAP estimate of w by setting the derivative with respect to $\hat{w}$ to zero. That gives the following equation to solve for $\hat{w}$.

$$\frac{y-\hat{w}}{\sigma_n^2} + f'(\hat{w}) = 0 \tag{10}$$

We now need a model $p_w(w)$ for the distribution of transform-domain coefficients, w. The pdf for wavelet coefficients of natural images $p_w(w)$ is often modeled as a generalized (heavy-tailed) Gaussian [6],

$$p_w(w) = K(s,p) \cdot \exp\left(-\left|\frac{w}{s}\right|^p\right) \tag{11}$$

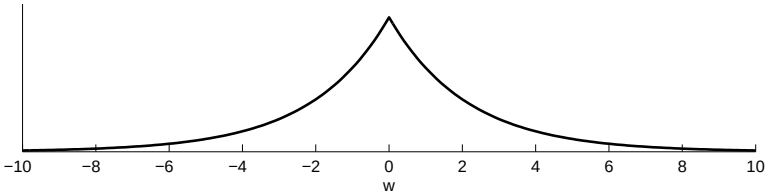
where s, p are the parameters for this model, and $K(s,p)$ is the normalization constant (which depends on s and p). Other pdf models have also been proposed [4, 5, 7, 3, 2].

We assume here that the coefficients can be modeled using a Laplacian pdf,

$$p_w(w) = \frac{1}{\sqrt{2}\,\sigma} \exp\left(-\frac{\sqrt{2}}{\sigma}\,|w|\right). \tag{12}$$

Even though the Laplace pdf may not the most accurate probability model for the distribution of the coefficients w, it is simple and at least captures the basic behaviour that is usually exhibited (heavier tails than the Gaussian density, and more peaked at the center than the Gaussian).

The pdf, $p_w(w)$, of a zero-mean Laplace random variable.



In this case

$$f(w) = -\log(\sigma\sqrt{2}) - \frac{\sqrt{2}}{\sigma} \cdot |w|$$

and so

$$f'(w) = -\frac{\sqrt{2}}{\sigma} \cdot \text{sign}\,(w).$$

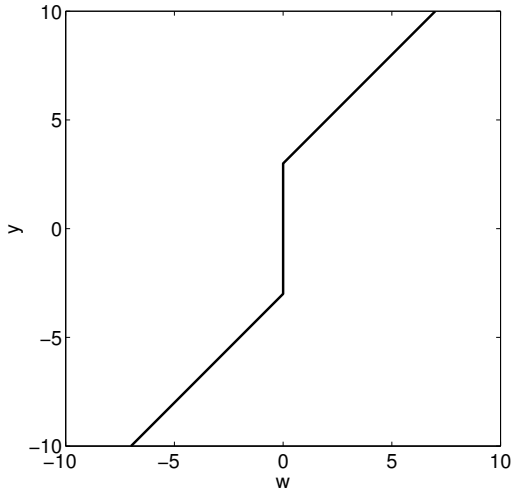


Figure 1: y as a function of $\hat{w}$ in (13).

Plugging into (10) gives

$$\frac{y-\hat{w}}{\sigma_n^2} - \frac{\sqrt{2}}{\sigma} \cdot \text{sign}\,(\hat{w}) = 0$$

or

$$y = \hat{w} + \frac{\sqrt{2}\,\sigma_n^2}{\sigma} \cdot \text{sign}\,(\hat{w}) \tag{13}$$

A graph of this relationship between y and $\hat{w}$ is illustrated in Figure 1. Therefore, a graph of $\hat{w}$ as a function of y, is illustrated in Figure 2. This graph is given by

$$\hat{w}(y) = \begin{cases} y+T, & y < -T \\ 0, & -T \le y \le T \\ y-T, & T < y \end{cases} \tag{14}$$

This is the *soft threshold* nonlinearity. The MAP estimate of w uses the threshold

$$T = \frac{\sqrt{2}\,\sigma_n^2}{\sigma}.$$

The formula (14) is often written as

$$\hat{w}(y) = \text{sign}(y) \cdot (|y| - T)_+ \tag{15}$$

where $(a)_+$ is defined as

$$(a)_+ = \begin{cases} 0 & \text{if } a < 0 \\ a & \text{if } a \ge 0. \end{cases} \tag{16}$$

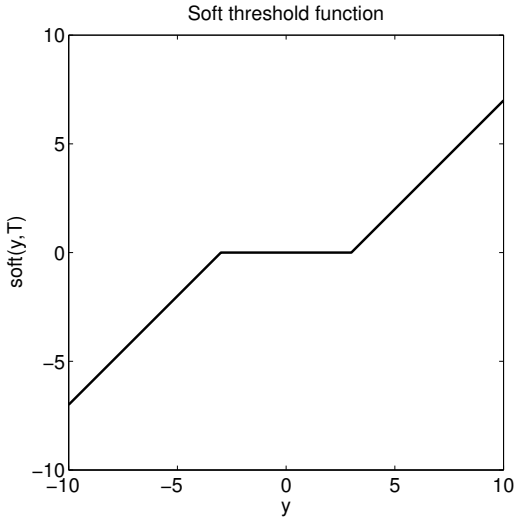


Figure 2: The soft-threshold function (14).

Lets define the soft operator as

$$\text{soft}(g, \tau) := \text{sign}(g) \cdot (|g| - \tau)_+ \tag{17}$$

then the MAP estimator (15) can be written as

$$\hat{w}(y) = \text{soft}\Big(y, \frac{\sqrt{2}\,\sigma_n^2}{\sigma}\Big). \tag{18}$$

2 Using the Soft-Threshold Function

To apply the soft-threshold rule we need to know σ_n and σ. Recall, σ_n is the standard deviation of the noise, and σ is the standard deviation of the noise-free transform coefficients. In the following, we assume that σ_n is known, but not σ.

Experiments with the wavelet transform show that σ is quite different from scale to scale, so we should estimate a different σ for each subband. The results is a *subband-dependent* threshold. For each subband, we must estimate σ from only the noisy data.

Because the transform-domain coefficients of the noise-free signal and the noise are independent, we have

$$\text{VAR}[y] = \text{VAR}[w] + \text{VAR}[n]$$

or

$$\text{VAR}[w] = \text{VAR}[y] - \text{VAR}[n].$$

As we assume we know the variance of the noise, we write

$$\sigma^2 = \text{VAR}[y] - \sigma_n^2.$$

The variance of y can be computed from each subband using the sample mean, where we assume all quantities are zero-mean,

$$\text{VAR}[y] = \text{MEAN}[y^2].$$

So we estimate σ as

$$\hat{\sigma} = \sqrt{\text{MEAN}[y^2] - \sigma_n^2}.$$

In case we have a negative value under the square root (it is possible because these are estimates) we can use

$$\hat{\sigma} = \sqrt{\max\left(\text{MEAN}[y^2] - \sigma_n^2, 0\right)}.$$

3 Remarks

1. Instead of the soft-threshold, other nonlinear shrinkage functions can be used. For example, instead of using the MAP estimator of a Laplace random variable in Gaussian noise, we can use the MMSE estimator. Other nonlinear shrinkage functions (not related to the Laplace probability model) are: hard-thresholding, garrot-threshold, and shrinkage functions derived from mixture models, etc. Some of these nonlinear shrinkage functions give better results than soft-thresholding.

2. Instead of processing each transform-domain coefficient individually, better denoising results can be achieved by processing groups of coefficients together. This requires a more complicated probability model for the coefficients. Specifically, it requires the joint probability density function of a group of coefficients.

3. The performance of transform-domain thresholding for noise reduction depends on the transform. For example, instead of the critically-sampled wavelet transform, other types of wavelet transforms can be used to obtain better denoising results. For example, the undecimated wavelet transform, the dual-tree complex wavelet transform, and curvelet transform can provide better results than the critically-sampled wavelet transform.

4 Exercises

1. Implement wavelet-based signal/image denoising or STFT-based speech denoising using the soft-threshold function,

2. Compare hard-thresholding and soft-thresholding for signal denoising.

3. Make up a new nonlinear threshold function of your own that is a compromise between soft and hard thresholding. Use it for signal/image denoising and compare it with the soft threshold (and compare it with hard thresholding, if you have implemented that).

4. Instead of the threshold

$$T = \sqrt{2}\,\frac{\sigma_n^2}{\sigma}$$

a different value is suggested in the paper [1]. Read the paper and find out what threshold value it suggests and why.

5. Suppose the noise-free coefficient w is modeled as zero-mean Gaussian instead of Laplace. (Assume still that the noise is zero-mean Gaussian.) In this case, the MAP estimate of w from the noisy data y is not the soft-threshold function, but a different function. Repeat the derivation of the MAP estimate to find out what the function is. How do you expect its performance for wavelet-based denoising to compare with the performance of the soft-threshold? Optional: perform signal/image denoising, but use the new function instead of the soft-threshold, and compare the result to the result of soft-thresholding.

6. We have assumed the use of an orthonormal transform. If the transform is not orthonormal, then the noise variance in the transform domain will be different than the noise variance in the signal domain. Suppose the noise variance in the signal domain is known and denoted σ_n^2. Find an expression for the noise variance in the transform domain in terms of the the transform matrix. That means, after the transform (eg, discrete wavelet transform or STFT) is performed on the noisy data, what is noise variance of the coefficients?

References

[1] S. G. Chang, B. Yu, and M. Vetterli. Adaptive wavelet thresholding for image denoising and compression. *IEEE Trans. on Image Processing*, 9(9):1532–1546, September 2000.

[2] M. A. T. Figueiredo and R. D. Nowak. Wavelet-based image estimation: An empirical Bayes approach using Jeffrey's noninformative prior. *IEEE Trans. on Image Processing*, 2001.

[3] H. Gao. Wavelet shrinkage denoising using the non-negative garrote. *Jour. of Comput. and Graph. Stat.*, 7:469–488, 1998.

[4] A. Hyvarinen. Sparse code shrinkage: Denoising of nongaussian data by maximum likelihood estimation. *Neural Computation*, 11:1739–1768, 1999.

[5] A. Hyvärinen, E. Oja, and P. Hoyer. Image denoising by sparse code shrinkage. In S. Haykin and B. Kosko, editors, *Intelligent Signal Processing*. IEEE Press, 2001.

[6] E. P. Simoncelli. Bayesian denoising of visual images in the wavelet domain. *Bayesian Inference in Wavelet Based Models. eds. P Müller and B Vidakovic. Springer-Verlag, Lecture Notes in Statistics 141*, March 1999.

[7] B. Vidakovic. *Statistical Modeling by Wavelets.* John Wiley & Sons, 1999.