Last Time: Decay of Fourier vs. Wavelet Coefficients

Toy Problem:

- $f: [0,1] \to \mathbb{R}$
- f is piecewise constant with S pieces
- $H = \sum_{k=1}^{S} |h_k|$ is the sum of the jumps

Fourier & Haar Wavelet Bases:

Fourier:

$$f(t) = \sum_{n \in \mathbb{Z}} \underbrace{\langle f, e^{j2\pi n t} \rangle}_{c_n} e^{j2\pi n t}$$

Wavelet:

$$f(t) = \int_0^1 f(t)\, dt + \sum_{i=0}^{\infty} \sum_{n=0}^{2^j - 1} \underbrace{\langle f, \psi_{i,n} \rangle}_{\theta_{i,n}} \psi_{i,n}(t)$$

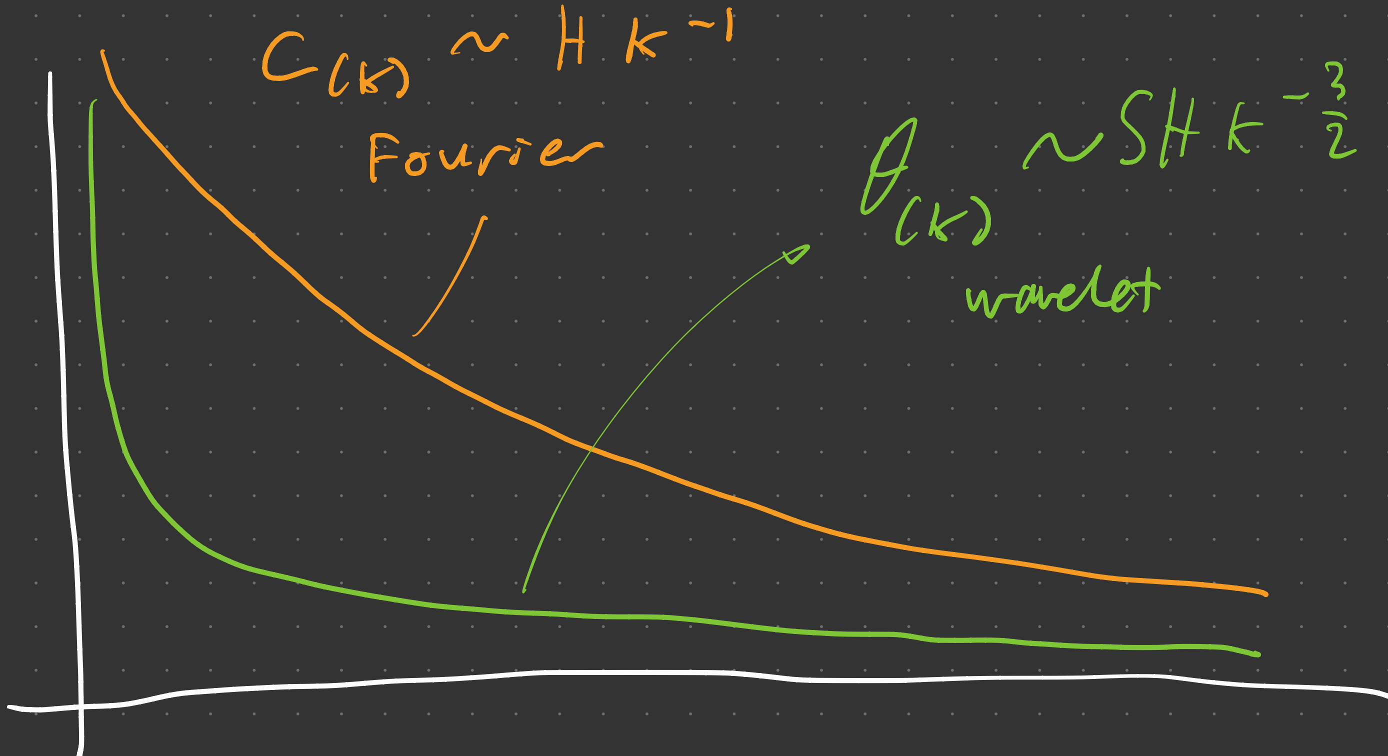
Let $c_{(k)}$ denote the kth largest Fourier coeff, i.e.,

$$|c_{(1)}| \geq |c_{(2)}| \geq |c_{(3)}| \geq \cdots$$

Let $\theta_{(k)}$ denote the kth largest wavelet coeff, i.e.,

$$|\theta_{(1)}| \geq |\theta_{(2)}| \geq |\theta_{(3)}| \geq \cdots$$

Sorted Fourier vs. Wavelet coeffs:



<u>Obs:</u> wavelet coefficients decay faster! sparsity

<u>Today:</u> Translate the decay rate to an approximation error rate.

Approximation in Bases

Q: Given any orthobasis $\{b_k\}_{k=1}^{\infty}$ of $L^2[0,1]$, how do we construct the best N-term approximation?

A: Threshold to only keep the N largest coefficients.

$$f(t) = \sum_{k=1}^{\infty} \underbrace{\langle f, b_k \rangle}_{a_k} b_k(t)$$

Let

$$|a_{(1)}| \geq |a_{(2)}| \geq |a_{(3)}| \geq \cdots$$

be a rearrangement of the $\{a_k\}_{k=1}^{\infty}$ in non-increasing order.

Obs: The best N-term approximation of f is

$$f_M(t) = \sum_{k=1}^{N} a_{(k)} b_{(k)}(t)$$

<u>Alternatively:</u> $f_M(t) = \sum_{k=1}^{\infty} \tau(a_k)\, b_k(t)$,

where

$$\tau(a) = \begin{cases} a, & \text{if } |a| > |a_{k+1}| \\ 0, & \text{else} \end{cases}$$

is the <u>hard-thresholding</u> operator.

<u>Q:</u> What is the approximation error?

$$\|f - f_M\|_{L^2}^2 = \int_0^1 |f(t) - f_M(t)|^2\, dt$$

$$= \int_0^1 \left| \sum_{k=1}^{\infty} a_{(k)} b_{(k)}(t) - \sum_{k=1}^{N} a_{(k)} b_{(k)}(t) \right|^2 dt$$

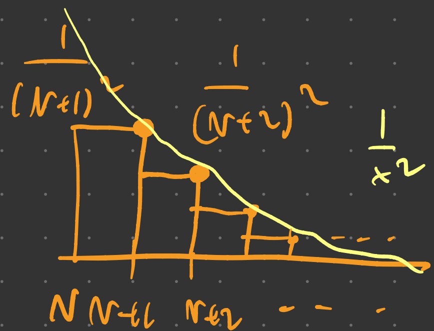
$$= \int_0^1 \left| \sum_{k=N+1}^{\infty} a_{(k)} b_{(k)}(t) \right|^2 dt$$

$$\leq \sum_{k=N+1}^{\infty} |a_{(k)}|^2 \underbrace{\int_0^1 |b_{(k)}(t)|^2\, dt}_{1}$$

<u>Obs:</u> This is the sum of the squares of the <u>tail</u> of the sorted coeffs.

Approximation Errors of Fourier vs. Wavelet

Fourier: $\|f - f_M^{\text{Fourier}}\|_{L^2}^2 \le C_F H \sum_{k=N+1}^{\infty} \frac{1}{k^2}$



$$\le C_F H \int_N^{\infty} \frac{1}{x^2}\, dx$$

$$= C_F H \left[-\frac{1}{x} \right|_N^{\infty}$$

$$= \frac{C_F H}{N}$$

Wavelet: $\|f - f_M^{\text{wavelet}}\|_{L^2}^2 \le C_W S H \sum_{k=N+1}^{\infty} \frac{1}{k^3}$

$$\le C_W S H \int_N^{\infty} \frac{1}{x^3}\, dx$$

$$= C_W S H \left[-\frac{1}{2} x^{-2} \right|_N^{\infty}$$

$$= \frac{C_W S H}{2N^2}$$

Summary:

For piecewise constant signals, the best N-term (squared) approx. error with

- Fourier decays as $\Theta(N^{-1})$
- Wavelet decays as $\Theta(N^{-2})$

and these rates are sharp.

Remark: Story is similar for piecewise poly. signals with higher-order wavelets.

- Story is also similar for signals with certain Besov regularity.

Denoising by Soft-Thresholding

Setup:

- $f : [0,1] \to \mathbb{R}$ is some analog signal

- observe

  Haar scaling func.

$$y_n = \langle f, \varphi_{I,n} \rangle + \varepsilon_n, \quad n = 0, \dots, 2^I - 1$$

$$\varepsilon_n \sim \mathcal{N}(0, \sigma^2) \text{ i.i.d.}$$

Goal: Denoise these measurements and recover $f(t)$

Algorithm:

- Compute an $(I-1)$-level DWT on $\{y_n\}_{n=0}^{2^I-1}$

$$\frac{1}{2^{I-1}} \sum_{n=0}^{2^I-1} y_n$$

$$\longrightarrow \tilde{f}(t) = \hat{c} + \sum_{i=1}^{I-1} \sum_{n=0}^{2^i-1} \tilde{\theta}_{i,n} \psi_{i,n}(t)$$

- Soft-threshold the coefficients $\tilde{\theta}_{i,n}$ with threshold level

  natural log

$$T \sim \underbrace{\sqrt{2 \cdot 2^{-I} \cdot I \cdot \log 2}}_{\lambda}$$

$$\tau(a) = \underbrace{\frac{a}{|a|}}_{\mathrm{sgn}(a)} \max\{|a| - \lambda, 0\}$$

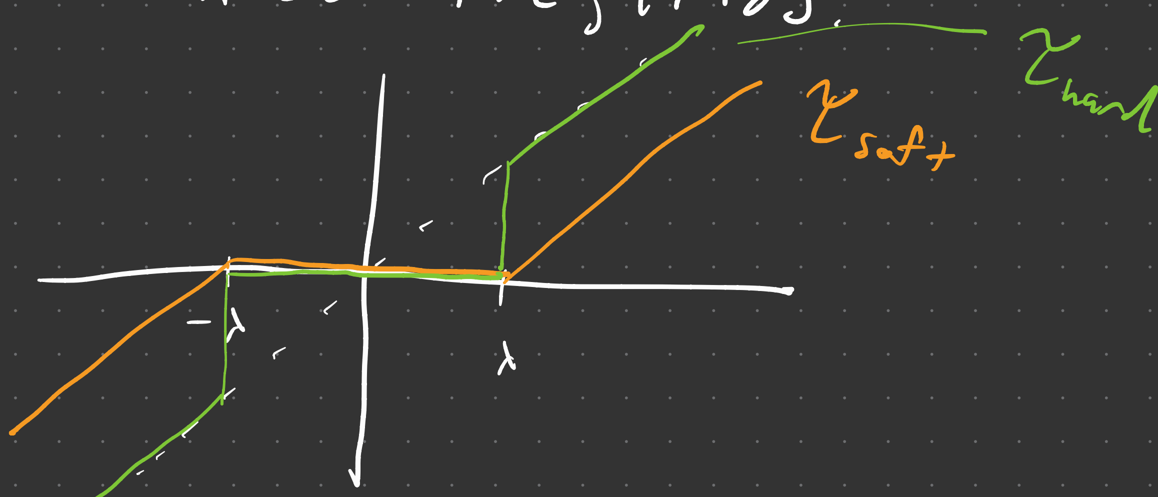
$$\longrightarrow \hat{\theta}_{i,n} = \tau(\tilde{\theta}_{i,n})$$

"denoising by soft-thresholding"

- Denoised signal

$$\hat{f}(t) = \hat{c} + \sum_{i=1}^{I-1} \sum_{n=0}^{2^i - 1} \hat{\theta}_{i,n} \psi_{i,n}(t)$$

<u>Remark</u>: $\hat{f}(t)$ is a <u>minimax</u> <u>optimal</u> estimator for $f(t)$ for most signal models. (Donoho & Johnstone, 1998)



Theorem (Donoho & Johnstone, 1998):

Let $\theta \in \mathbb{R}^p$ be a vector. Suppose we observe

$$Y_k = \theta_k + \varepsilon_k$$

where $\varepsilon_k \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. Let

$$\hat{\theta}_k = \frac{y_k}{|y_k|} \max\{|y_k| - \lambda, 0\}.$$

soft thresholding

Then

$$\mathbb{E}\left[\|\theta - \hat{\theta}\|_2^2\right] \leq (2\log p + 1)\left(\sigma^2 + \sum_{k=1}^{p} \min\{\theta_k^2, \sigma^2\}\right)$$

mean-squared error

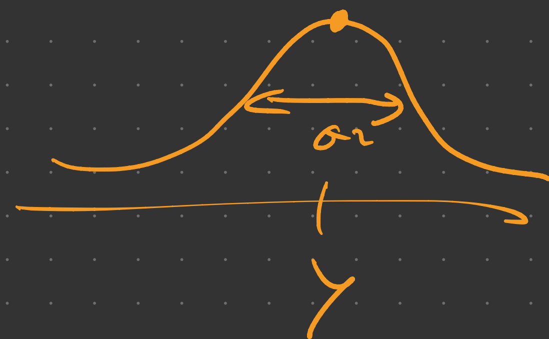
Intuition:

$$y = \theta + \varepsilon \in \mathbb{R}^p, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

Q: What is the Maximum-liklihood estimator?

A: $\text{MLE}(\theta) = y$

$$\left[\theta \sim \mathcal{N}(y, \sigma^2 I)\right]$$



MSE of MLE:

$$E\left[\|\theta - \text{MLE}(\theta)\|_2^2\right] = p\sigma^2$$

Q: What if we knew θ had only s nonzero coeffs AND we knew where they were?

A: We would only need to estimate those coeffs:

$$\text{MSE} = s\,\sigma^2 \ll p\sigma^2$$

Remark: soft-thresholding allows us to do better than the MLE when θ is sparse or approximately sparse.

Consider the "oracle estimator"

$$\hat{\theta}_k^{O} = \begin{cases} Y_k, & |\theta_k|^2 \geq \sigma^2 \\ 0, & |\theta_k|^2 < \sigma^2 \end{cases}$$

only estimate if signal power exceeds noise power

$\rightarrow$ not realizable since we don't know θ_k.

$$MSE = \mathbb{E}\left[\sum_{k=1}^{p} |\hat{\theta}_k^{O} - \theta_k|^2\right]$$

$$= \sum_{k=1}^{p} \min\{|\theta_k|^2, \sigma^2\}$$

<u>Obs:</u> Up to a log factor, the soft-thresholding estimator is as good as the oracle estimator.

Suppose θ has s nonzero coeffs with size $\geq \sigma$. Then,

- Oracle MSE $= s\sigma^2$
- soft-threshold MSE $= (2\log p + 1)(s+1)\sigma^2$

$$\approx 2s \log p \,\sigma^2$$

Remark:

$$\hat{\theta} = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} \|y - \theta\|_2^2 + \lambda \|\theta\|_1$$

$$L_1 = \sum_{k=1}^{p} |\theta_k|$$

- compressed sensing
- ℓ^1-minimization algorithms
- etc.

Analysis of Soft-Thresholding

Let $\boldsymbol{\theta} \in \mathbb{R}^p$ be a vector of coefficients/parameters. Suppose we observe

$$\boldsymbol{y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}\ ,\ \boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})\ .$$

The MLE of $\boldsymbol{\theta}$ is simply $\boldsymbol{y}$, and its mean-square error is

$$\mathbb{E}\|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2 \ = \ p\sigma^2$$

However, suppose that only k of the coefficients are nonzero. If we knew which k these were, then we would only need to estimate those. The resulting estimator $\widehat{\boldsymbol{\theta}}$, which sets all but the k coefficients to zero, would have

$$\mathbb{E}\|\boldsymbol{y} - \boldsymbol{\theta}\|_2^2 \ = \ k\sigma^2$$

Of course, in practice we would not know which coefficients were zero. The soft-thresholding estimator is a data-based way of deciding which coefficients should be estimated to be zero.

$$\widehat{\theta}_i = \text{sign}(y_i)\ \max(|y_i| - \lambda, 0)\ ,\ \lambda > 0$$

This can perform much better than the MLE if $\boldsymbol{\theta}$ is sparse or approximately sparse.

Before we analyze the soft-thresholding estimator, let us consider an ideal thresholding estimator. Suppose that an oracale tells us the magnitude of each θ_i. The *oracle* estimator is

$$\widehat{\theta}_i^O \ = \ \begin{cases} y_i & \text{if } |\theta_i|^2 \geq \sigma^2 \\ 0 & \text{if } |\theta_i|^2 < \sigma^2 \end{cases}$$

In other words, we estimate a coefficient if and only if the signal power is at least as large as the noise power. The MSE of this estimator is

$$\mathbb{E}\sum_{i=1}^p |\widehat{\theta}_i^O - \theta_i|^2 = \sum_{i=1}^p \min(|\theta_i|^2, \sigma^2)$$

Notice that the MSE of the oracle estimator is always less than or equal to the MSE of the MLE. If $\boldsymbol{\theta}$ is sparse, then the MSE of the oracle estimator can be much smaller. If all but $k < p$ coefficients are zero, then the MSE of the oracle estimator is at most $k\sigma^2$. Remarkably, the soft-thresholding estimator comes very close to achieving the performance of the oracle, as shown by the following theorem (Theorem 1 in "Ideal Spatial Adaptation by Wavelet Thresholding," by Donoho and Johnstone).

The theorem uses the threshold $\lambda = \sqrt{2\sigma^2 \log p}$. This choice of threshold is motivated by the following observation. Assume, for the moment, that all the coefficients are zero (i.e., $\theta_i = 0$ for $i = 1, \ldots, p$). In this case, we should set the threshold so that it is larger than the magnitude of any of the y_i (so they are all set to zero). If we take $\lambda = \sqrt{2\sigma^2 \log \frac{p}{\delta}}$, then using the Gaussian tail bound and the union bound we have $\mathbb{P}(\bigcup_{i=1}^p \{|y_i| \geq \lambda\}) \leq \delta$.

Theorem 1. *Assume the direct observation model above and let*

$$\widehat{\theta}_i = \text{sign}(y_i)\ \max(|y_i| - \lambda, 0)$$

with $\lambda = \sqrt{2\sigma^2 \log p}$*. Then*

$$\mathbb{E}\|\widehat{\theta} - \theta\|_2^2 \ \le\ (2\log p + 1)\left\{\sigma^2 + \sum_{i=1}^{p} \min(|\theta_i|^2, \sigma^2)\right\}$$

The theorem shows that the soft-thresholding estimator mimics the MSE performance of the oracle estimator to within a factor of roughly $2\log p$. For example, if $\boldsymbol{\theta}$ is k-sparse (with non-zero coefficients larger than σ in magnitude), then the MSE of the oracle is $k\sigma^2$ and the MSE of the soft-thresholding estimator is at most $(2\log p + 1)(k+1)\sigma^2 \approx 2k\log p\, \sigma^2$ when n is large. This also corresponds to a huge improvement over the MLE if $2k\log p \ll p$.

Intution: Consider the case with $\sigma^2 = 1$ (the general case follows by simple rescaling). First recall that if $y \sim \mathcal{N}(0,1)$, then $\mathbb{P}(|y| \ge \lambda) \le e^{-\lambda^2/2}$. This inequality is easily derived as follows. Since $\mathbb{P}(y \ge \lambda) = \mathbb{P}(y \le -\lambda)$, we only need to show that $\mathbb{P}(y \ge \lambda) = \frac{1}{2\pi}\int_\lambda^\infty e^{-y^2/2}dy \le \frac{1}{2}e^{-\lambda^2/2}$. Note that

$$\frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-y^2/2}dy}{\frac{1}{2}e^{-\lambda^2/2}} \ = \ \frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-(y^2-\lambda^2)/2}dy}{\frac{1}{2}} \ = \ \frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-(y-\lambda)(y+\lambda)/2}dy}{\frac{1}{2}}\ .$$

The desired inequality results by making change of variable $t = y + \lambda$ to yield

$$\frac{\frac{1}{2\pi}\int_\lambda^\infty e^{-y^2/2}dy}{\frac{1}{2}e^{-\lambda^2/2}} \ = \ \frac{\frac{1}{2\pi}\int_0^\infty e^{-t(t+2\lambda)/2}dt}{\frac{1}{2}} \ \le \ \frac{\frac{1}{2\pi}\int_0^\infty e^{-t^2/2}dt}{\frac{1}{2}} \ = \ 1\ .$$

Now observe that if $\lambda = \sqrt{2\log p}$, then $\mathbb{P}(|y_i| \ge \lambda | \theta_i = 0) \le e^{-\log p} = \frac{1}{p}$. Using this we have

$$\mathbb{E}\left[\sum_{i:\theta_i=0} \mathbb{1}\left\{\widehat{\theta}_i \neq 0\right\}\right] \ = \ \sum_{i:\theta_i=0} \frac{1}{p} \ \le \ 1\ .$$

In other words, using this threshold we expect that at most one of the $\theta_i = 0$ will not be estimated as $\widehat{\theta}_i = 0$. Next consider cases when $\theta_i \neq 0$. Let's suppose that $|\theta_i| \gg \lambda$, so that $\widehat{\theta}_i = y_i - \lambda\text{sign}(y_i)$. In this case,

$$(\theta_i - \widehat{\theta}_i)^2 \ = \ (-\epsilon_i + \lambda\text{sign}(y_i))^2 \ \le \ \epsilon_i^2 + 2|\epsilon_i|\lambda + \lambda^2\ .$$

Taking the expecation of this upper bound yields

$$\mathbb{E}[(\theta_i - \widehat{\theta}_i)^2] \ \le \ 1 + 2\lambda + \lambda^2 \ \le \ 3\lambda^2 + 1\ , \text{ assuming } \lambda > 1\ .$$

Thus, if $\boldsymbol{\theta}$ has only k nonzero weights, then this intution suggests that

$$\sum_{i=1}^{p} \mathbb{E}[(\theta_i - \widehat{\theta}_i)^2] \ = \ O(k\log p)\ .$$

This is formalized in the following proof of Theorem 1.

Proof: To simplify the analysis, assume that $\sigma^2 = 1$. The general result follows directly. It suffice to show that

$$\mathbb{E}[(\widehat{\theta}_i - \theta_i)^2] \;\leq\; (2\log p + 1)\left\{\frac{1}{p} + \min(\theta_i^2, 1)\right\}$$

for each i. So let $y \sim \mathcal{N}(\theta, 1)$ and let $f_\lambda(y) = \text{sign}(y)\max(|y| - \lambda, 0)$. We will show that with $\lambda = \sqrt{2\log p}$

$$\mathbb{E}[(f_\lambda(y) - \theta)^2] \;\leq\; (2\log p + 1)\left\{\frac{1}{p} + \min(\theta^2, 1)\right\}\;.$$

First note that $f_\lambda(y) = y - \text{sign}(y)(|y| \wedge \lambda)$, where $a \wedge b$ is shorthand notation for $\min(a, b)$. It follows that

$$\begin{aligned}
\mathbb{E}[(f_\lambda(y) - \theta)^2] &= \mathbb{E}[(y-\theta)^2] - 2\mathbb{E}[\text{sign}(y)(|y| \wedge \lambda)(y-\theta)] + \mathbb{E}[y^2 \wedge \lambda^2] \\
&= 1 - 2\mathbb{E}[\text{sign}(y)(|y| \wedge \lambda)(y-\theta)] + \mathbb{E}[y^2 \wedge \lambda^2]
\end{aligned}$$

The expected value in the second term is equal to $\mathbb{P}(|y| < \lambda)$, which is verified as follows.

The expectation can be split into integrals over four intervals, $(\infty, -t]$, $(-t, 0]$, $(0, t]$, and (t, ∞). Each integrand is a linear or quadratic function of y times the Gaussian density function. Let $\phi(x) := \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$ and $\Phi(x)$ be the cumulative distribution function of $\phi(x)$, and consider the following indefinite Gaussian integral forms:

$$\begin{aligned}
\int \phi(x)\,dx &= \Phi(x)\;,\ \text{by definition of } \Phi, \\
\int x\phi(x)\,dx &= \frac{1}{\sqrt{2\pi}}\int xe^{-x^2/2}\,dx \;=\; \underbrace{-\frac{1}{\sqrt{2\pi}}\int e^u\,du}_{u=-x^2/2} \;=\; -\frac{1}{\sqrt{2\pi}}e^u \;=\; -\phi(x)\;, \\
\int x^2\phi(x)\,dx &= \Phi(x) - x\phi(x)\;.
\end{aligned}$$

The last form is verified as follows. Let $u = x$ and $dv = x\phi(x)dx$. Then integration by parts $\int u dv = uv - \int v du$ and $\int x\phi(x)dx = -\phi(x)$ show that

$$\int x^2\phi(x)\,dx = x\int x\phi(x)dx - \int\int x\phi(x)dx = -x\phi(x) + \int \phi(x) = \Phi(x) - x\phi(x)\;.$$

The Gaussian distribution we are considering has mean θ so the shifted integral forms below, which follow immediately from the derviations above by variable substitution, will be used in our analysis:

$$\begin{aligned}
(i) \quad & \int \phi(x-\theta)dx &&= \Phi(x-\theta) \\
(ii) \quad & \int x\phi(x-\theta)dx &&= \theta\Phi(x-\theta) - \phi(x-\theta) \\
(iii) \quad & \int x^2\phi(x-\theta)dx &&= (1+\theta^2)\Phi(x-\theta) - (x+\theta)\phi(x-\theta)
\end{aligned}$$

Using these forms we compute

$$\begin{aligned}
\mathbb{E}[\text{sign}(x)(|x| \wedge \lambda)(x-\theta)] &= \int_{-\infty}^{\infty} \text{sign}(x)(|x| \wedge \lambda)(x-\theta)\,\phi(x-\theta)\,dx \\
&= \underbrace{\int_{-\infty}^{-\lambda} -\lambda(x-\theta)\phi(x-\theta)\,dx}_{\lambda\phi(-\lambda-\theta)} \underbrace{- \int_{-\lambda}^{0} x(x-\theta)\phi(x-\theta)dx}_{\Phi(-\theta)-\Phi(-\lambda-\theta)-\lambda\phi(-\lambda-\theta)} \\
&\quad + \underbrace{\int_{0}^{\lambda} x(x-\theta)\phi(x-\theta)dx}_{\Phi(\lambda-\theta)-\Phi(-\theta)-\lambda\phi(\lambda-\theta)} + \underbrace{\int_{\lambda}^{\infty} \lambda(x-\theta)\phi(x-\theta)dx}_{\lambda\phi(\lambda-\theta)} \\
&= \Phi(\lambda-\theta) - \Phi(-\lambda-\theta) = \mathbb{P}(|x| < \lambda)
\end{aligned}$$

So we have shown that

$$\mathbb{E}[(f_\lambda(x)-\theta)^2] = 1 - 2\mathbb{P}(|x| < \lambda) + \mathbb{E}[x^2 \wedge \lambda^2]$$

Note first that since $x^2 \wedge \lambda^2 \le \lambda^2$ we have

$$\mathbb{E}[(f_\lambda(x)-\theta)^2] \le 1 + \lambda^2 = 1 + 2\log p < (2\log p + 1)(1/p + 1)\ .$$

On the other hand, since $x^2 \wedge \lambda^2 \le x^2$ we also have

$$\mathbb{E}[(f_\lambda(x)-\theta)^2] \le 1 - 2\mathbb{P}(|x| < \lambda) + \theta^2 + 1 = 2(1 - \mathbb{P}(|x| < \lambda)) + \theta^2 = 2\mathbb{P}(|x| \ge \lambda) + \theta^2\ .$$

The proof will be finished if we show that

$$2\mathbb{P}(|x| \ge \lambda) \le (2\log p + 1)/p + (2\log p)\theta^2\ .$$

Define $g(\theta) := 2\mathbb{P}(|x| \ge \lambda)$ and note that g is symmetric about 0. Using a Taylor's series with remainder we have

$$g(\theta) \le g(0) + \frac{1}{2}\sup|g''|\theta^2\ ,$$

where g'' is the second derivative of g. Note that $g(\theta) = 2\,[1 - \mathbb{P}(z \le \lambda - \theta) + \mathbb{P}(z \le -\lambda - \theta)]$, where $z \sim \mathcal{N}(0,1)$. Using the Gaussian tail bound $\mathbb{P}(z > \lambda) \le \frac{1}{2}e^{-\lambda^2/2}$ and plugging in $\lambda = \sqrt{2\log p}$ we obtain $g(0) \le 2/p$. Note that $g'(\theta) = 2[\phi(\lambda-\theta) - \phi(-\lambda-\theta)]$ and $g'(0) = 0$. The integral (ii) above shows that the derivative of $\phi(\lambda-\theta)$ with respect to θ is equal to $(\lambda-\theta)\phi(\lambda-\theta)$. So we have $g''(\theta) = 2[(\lambda-\theta)\phi(\lambda-\theta) + (-\lambda-\theta)\phi(-\lambda-\theta)]$. It is easy to verify that $|g''(\theta)| < 1$. To simplify the final bound, note that $4\log p > 1$ if $p \ge 2$, so it follows that $\sup_\theta g''(\theta) < 4\log p$ for all $p \ge 2$. □