HW 5

Evan Ott
UT EID: eao466

October 10, 2016

Penalized likelihood and soft thresholding

(A)

First, show that $S_\lambda(y) = \arg\min_\theta \frac{1}{2}(y-\theta)^2 + \lambda|\theta|$ has the negative log-likelihood of a Gaussian distribution with mean θ and variance 1 as its quadratic term:

$$\begin{aligned} L(\theta|y) &= \frac{1}{\sqrt{2\pi \cdot 1}} \exp\left(-\frac{(y-\theta)^2}{2 \cdot 1}\right) \\ \log L(\theta|y) &= \log\left[\frac{1}{\sqrt{2\pi}}\right] - \frac{(y-\theta)^2}{2} = c - \frac{(y-\theta)^2}{2} \end{aligned}$$

So the negative log-likelihood is $\frac{(y-\theta)^2}{2} + c'$ for some $c' \in \mathbb{R}$ that does depends on neither y nor θ, which is exactly the quadratic term in $S_\lambda(y)$.

Now, let's prove the value

$$\begin{aligned} & S_\lambda(y) = \arg\min_\theta \frac{1}{2}(y-\theta)^2 + \lambda|\theta| \\ \theta > 0: \quad & S_\lambda(y) = \arg\min_\theta \frac{1}{2}(\theta-y)^2 + \lambda\theta \\ & \Rightarrow \quad 0 = \hat{\theta} - y + \lambda \\ & S_\lambda(y) = \hat{\theta} = y - \lambda \\ \theta < 0: \quad & S_\lambda(y) = \arg\min_\theta \frac{1}{2}(\theta-y)^2 - \lambda\theta \\ & \Rightarrow \quad 0 = \hat{\theta} - y - \lambda \\ & S_\lambda(y) = \hat{\theta} = y + \lambda \\ \theta = 0: \quad & S_\lambda(y) = 0 \end{aligned}$$

So, we can now look at the arg min as selecting one of those three values of $\hat{\theta}$, depending on which one has the smallest value of the objective function $f(\hat{\theta}) = \frac{1}{2}(y-\hat{\theta})^2 + \lambda|\hat{\theta}|$:

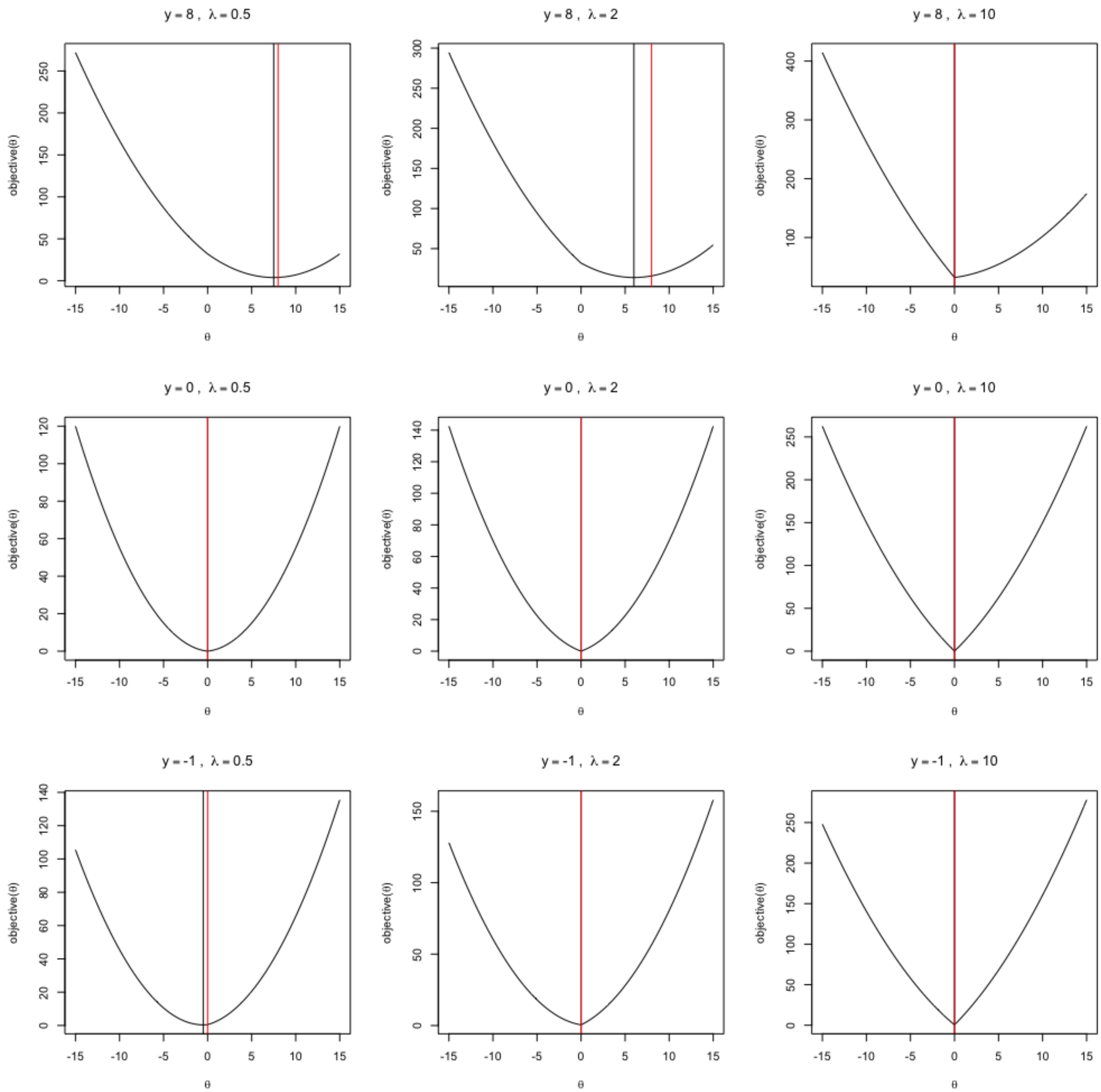
$$\begin{aligned} \theta > 0: \quad f(\hat{\theta}) &= \frac{1}{2}(y - [y-\lambda])^2 + \lambda|y-\lambda| \\ &= \frac{\lambda^2}{2} + \lambda|y-\lambda| \\ \theta < 0: \quad f(\hat{\theta}) &= \frac{1}{2}(y - [y+\lambda])^2 + \lambda|y+\lambda| \\ &= \frac{\lambda^2}{2} + \lambda|y+\lambda| \\ \theta = 0: \quad f(\hat{\theta}) &= \frac{1}{2}(y-0)^2 + \lambda|0| \\ &= \frac{y^2}{2} \end{aligned}$$

Because the second term is a penalty, we can assume $\lambda \geq 0$, whereas $y \in \mathbb{R}$. Thus, if $|y| \leq \lambda$, $\hat{\theta} = 0$ produces the smallest value of the objective function ($\hat{\theta} \neq 0$ has a quadratic term that's larger and adds a non-negative value to that). So $|y| - \lambda \leq 0$ produces $S_\lambda(y) = 0$ (in particular, $S_\lambda(y) = 0$).

Now, if $|y| > \lambda$ we must decide if $\hat{\theta} > 0$ or $\hat{\theta} < 0$. The quadratic terms are identical, as is the multiplier of the absolute value term, so the absolute value term is all we need to consider. Because $\lambda \geq 0$, if $y > 0$ then

$|y-\lambda| < |y+\lambda|$, so $S_\lambda(y) = y - \lambda = |y| - \lambda$. Similarly, if $y < 0$ then $|y-\lambda| > |y+\lambda|$ so $S_\lambda(y) = y + \lambda = -|y| + \lambda$. Therefore,

$$S_\lambda(y) = \begin{cases} 0 & y = 0 \\ 0 & |y| \leq \lambda \wedge y \neq 0 \\ |y| - \lambda & y > 0 \wedge |y| > \lambda \\ -(|y| - \lambda) & y < 0 \wedge |y| > \lambda \end{cases} = \text{sign}(y)(|y| - \lambda)_+$$



Notes from class

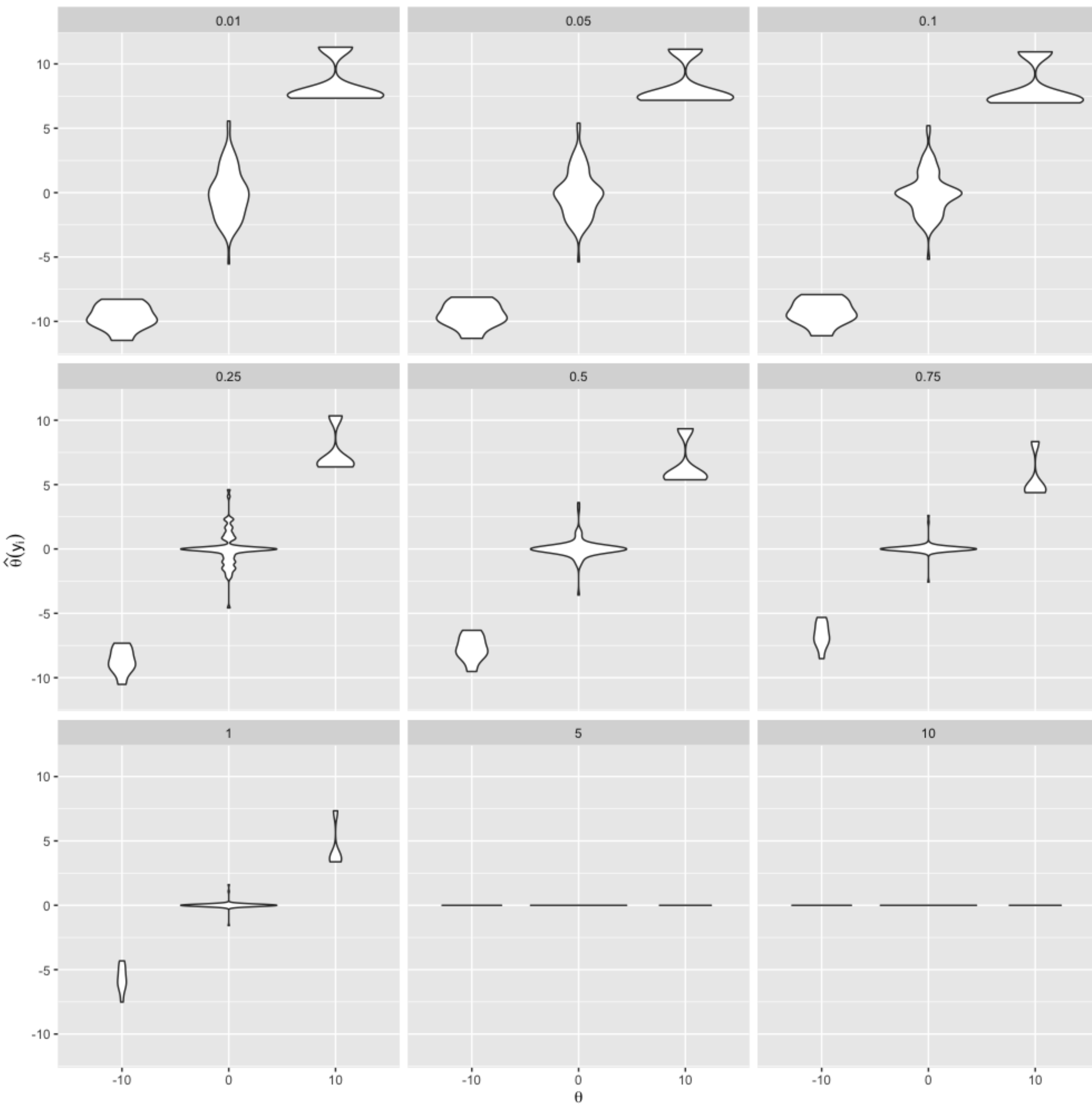
Subdifferential calculus handles things like $|x|$ by allowing more generality. If $f(x)$ has a gradient at x_0, then the only subgradient is the gradient, so the subdifferential is $\{\nabla f(x_0)\}$. If it doesn't have a gradient there (there's some sort of cusp), then we want $f(x) \geq f(x_0) + g \cdot (x - x_0)$ for all x (this restriction is harsh, and pretty much only works for convex functions). For example, if $f(x) = |x|$, then at $x_0 = 0$ $g = 0.5$ would satisfy this (by being below $|x|$ on either side of x_0). In this case, any $g \in [-1, 1]$ is a subgradient and $[-1, 1]$ is the subdifferential.

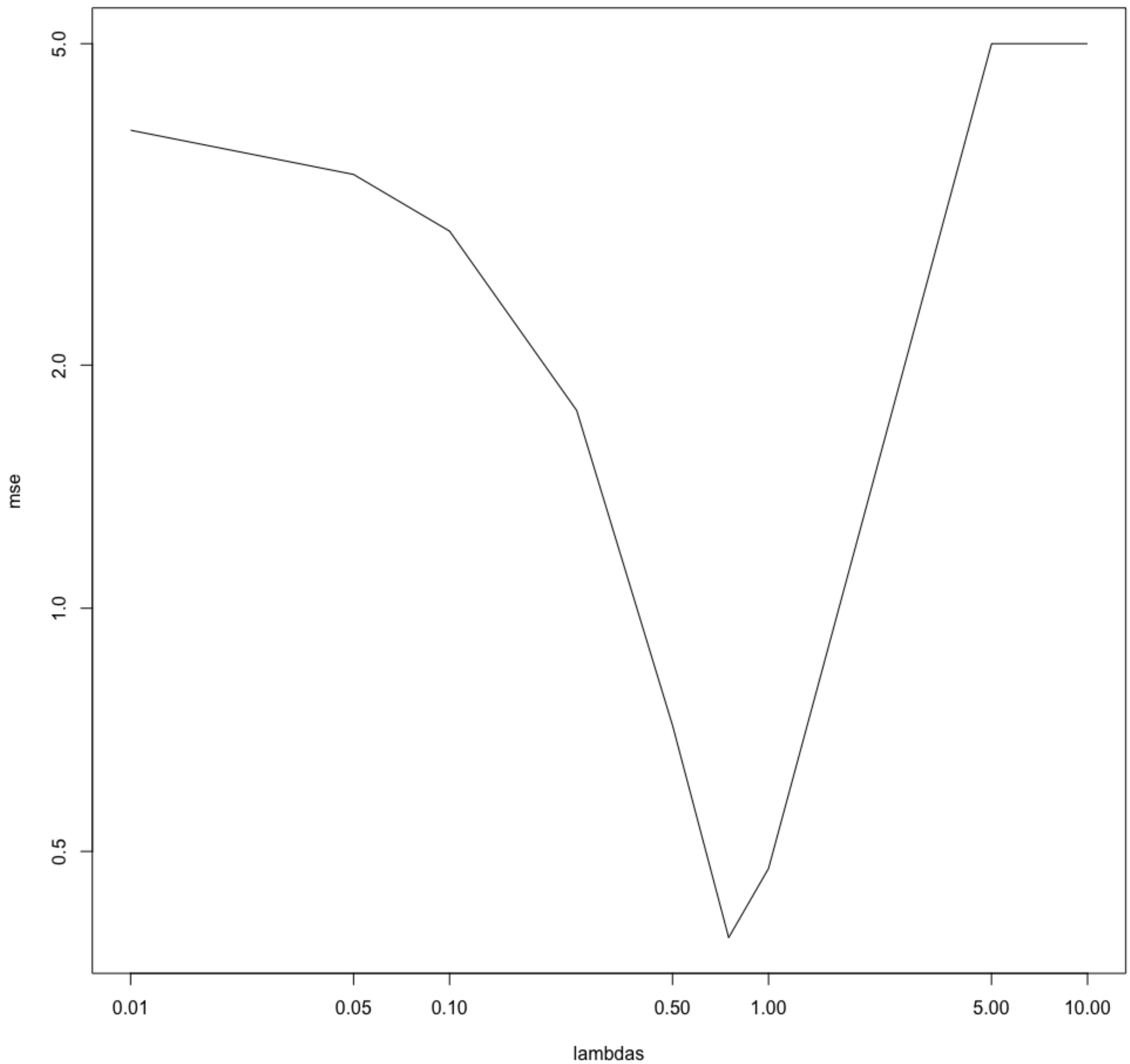
Theorem. Let $\partial f(x_0)$ be the subdifferential at a given x_0. If f is convex, then we have $0 \in \partial f(x_0) \Rightarrow x_0$ is optimal.

For example, $f(x) = |x|$ has

$$\partial f(x) = \begin{cases} \{\text{sign}(x)\} & x \neq 0 \\ [-1, 1] & x = 0 \end{cases}$$

(B)





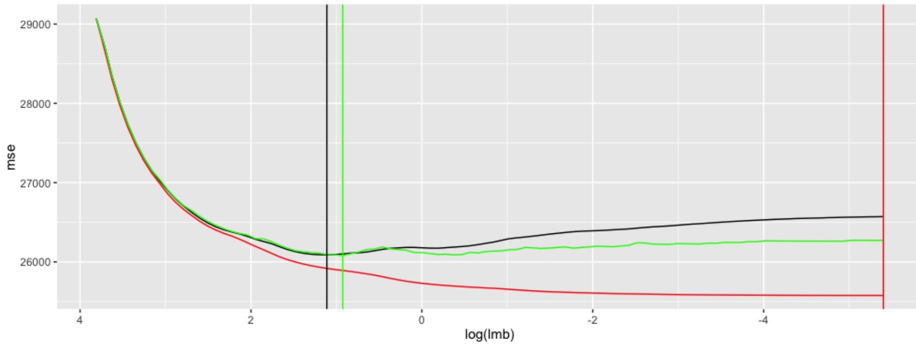
Notes from class

This independent normal means toy problem is useful for orthogonal decompositions for Fourier analysis or wavelets. The context of wavelets is:

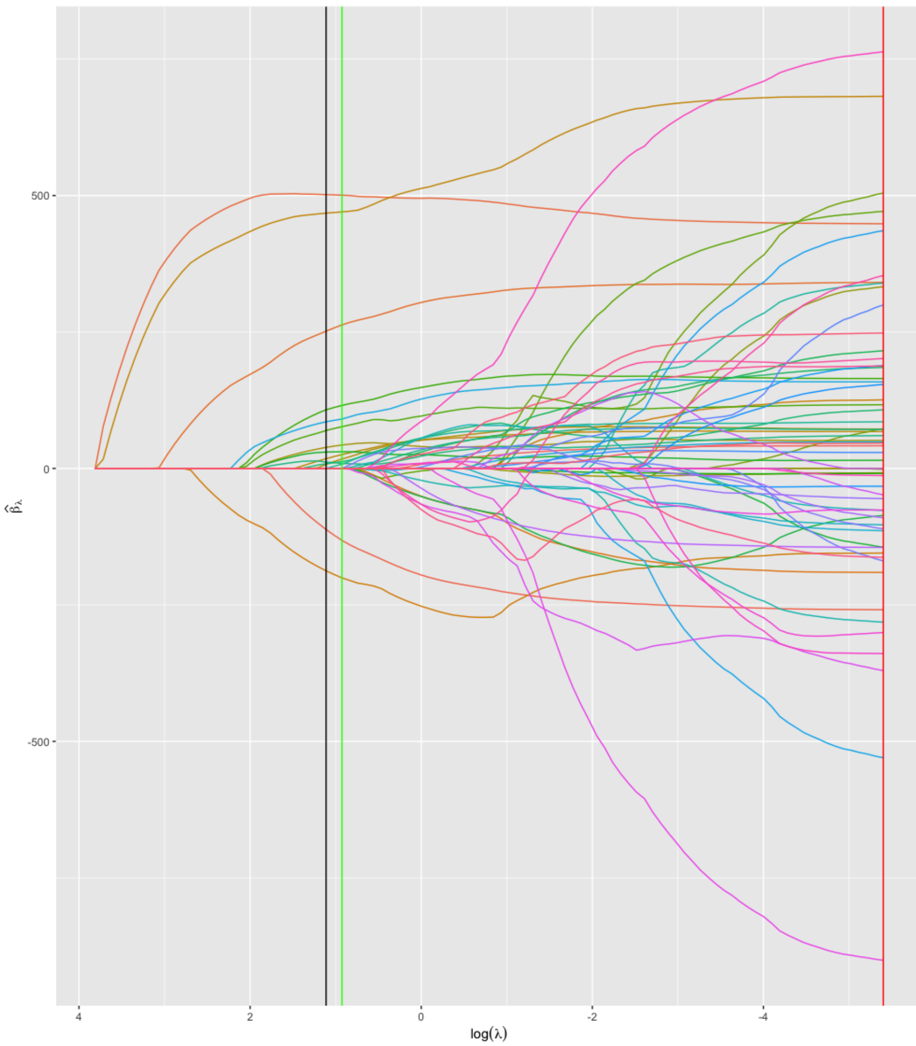
$$y_i = f(x_i) + e_i$$
$$f(x)* = \sum_{k=1}^{D} \theta_k \psi_k(x)$$

where ψ_k are basis functions and θ_k are coefficients. ψ_k in wavelets are generally not closed-form, but easy to graph (see Haar and Daubechies). This reduces down to a regression of normal means (especially if you have $n = 2^d$ observations for some d, then split the Haar basis functions d times and get 2^d values, that means that you get $D = 2^d$ and you don't want overfitting, so you're just fitting the θ_k by shrinking normal means.

The lasso



29000
28000
27000
26000
mse
4
2
0
-2
-4
log(lmb)



500
0
-500
$\hat{\beta}_\lambda$
4
2
0
-2
-4
log(λ)