Homework 3 Solution

1. (a)

$$\begin{aligned}\mathbb{E}(\hat{\theta}_j - theta_j) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} m(x_i)\psi_j(x_i) + \epsilon_i\psi_j(x_i) - \theta_j\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} m(x_i)\psi_j(x_i) - \int_0^1 m(x_i)\psi_j(x_i)dx\end{aligned}$$

Where we used linearity of expectation and the fact that ϵ_i is the only random quantity, and it has mean 0.

We can lower bound the integral by breaking the region into n blocks, each of length $1/n$, and replacing each term with the minimum value that that term takes on its interval, multiplied by the width of the interval. We'll upper bound the first term by taking the max over each interval. I.e. we can bound the last expression by:

$$\frac{1}{n}\sum_{i=1}^{n} m(x_i)\psi_j(x_i) - \int_0^1 m(x_i)\psi_j(x_i)dx \leq \frac{1}{n}\sum_{i=1}^{n}\left(\max_{\frac{i-1}{n}\leq z\leq\frac{i}{n}} m(z)\psi_j(z) - \min_{\frac{i-1}{n}\leq y\leq\frac{i}{n}} m(y)\psi_j(y)\right)$$

For brevity suppose that z_i achieves the maximum on the ith interval and y_i achieves the minimum on the ith interval. Then we are interested in bounding $m(z_i)\psi_j(z_i) - m(y_i)\psi_j(y_i)$. First, we can add and subtract $m(z_i)\psi_j(y_i)$ to get:

$$\begin{aligned}&m(z_i)\psi_j(z_i) - m(z_i)\psi_j(y_i) + m(z_i)\psi_j(y_i) - m(y_i)\psi_j(y_i) \\ &= m(z_i)\left(\psi_j(z_i) - \psi_j(y_i)\right) + \psi_j(y_i)\left(m(z_i) - m(y_i)\right)\end{aligned}$$

Our goal is to derive a bound for this expression, and we want to show that this expression decays at the rate $O(\frac{1}{\sqrt{n}})$, which will give us the correct rate at the end. First, we notice that $\psi_j(y_i)$ is upper bounded by $\sqrt{2}$. Moreover, using Cauchy-Schwarz, we can show that $m(z_i)$ is upper bounded:

$$\begin{aligned}m(z_i) &= \sum_{j=1}^{\infty}\theta_j\psi_j(z_i) \leq \sqrt{2}\sum_{j=1}^{\infty}\theta_j = \sqrt{2}\sum_{j=1}^{\infty}\frac{\theta_j j^\beta}{j^\beta} \\ &\leq \sqrt{2}\sqrt{\sum_{j=1}^{\infty}\theta_j^2 j^{2\beta}\sum_{j=1}^{\infty}\frac{1}{j^{2\beta}}} \\ &\leq \sqrt{\frac{C\pi^2}{3}}\end{aligned}$$

Where in the first line, we used that each $\psi_j(z_i)$ is upper bounded by $\sqrt{2}$. In the second line we just used Cauchy-Schwarz and in the third line, we used the condition placed on Θ and also that $\sum_{j=1}^{\infty}\frac{1}{j^2} = \frac{\pi^2}{6}$ and that the sum decreases with increasing β, meaning that $\frac{\pi^2}{6}$ is an upper bound for that sum.

Next, we need to bound the two differences. First we look at $\psi_j(z_i) - \psi_j(y_i)$. We use the fact that $\cos(x)$ has bounded derivative (i.e. $\sin(x)$ is bounded by 1), meaning that $|cos(x) - cos(y)| \leq C|x - y|$. In other words, we'll show that ψ_j is Lipschitz continuous:

$$\frac{\partial \psi_j(x)}{\partial x} = \sqrt{2}\pi j \sin(\pi j x) \leq \sqrt{2}\pi j$$
$$\Rightarrow |\psi_j(z_i) - \psi_j(y_i)| \leq \sqrt{2}\pi j |z_i - y_i|$$

Because the slope of the function is upper bounded by $\sqrt{2}\pi j$, so there is no way for the function to rise by more than that over the interval $|z_i - y_i|$. In particular, this condition for us means that:

$$\psi_j(z_i) - \psi_j(y_i) \leq \frac{\sqrt{2}\pi j}{n}$$

Next we need to bound $m(z_i) - m(y_i)$.

$$\begin{aligned} m(z_i) - m(y_i) &= \sum_{j=1}^{\infty} \theta_j(\psi_j(z_i) - \psi_j(y_i)) \\ &\leq \sum_{j=1}^{\infty} \theta_j \min(2, \sqrt{2}\pi \frac{j}{n}) \\ &\leq \sum_{j=1}^{N} \theta_j \sqrt{2}\pi \frac{j}{n} + \sum_{j=N+1}^{\infty} 2\theta_j \end{aligned}$$

where N is the smallest integer such that $2 \leq \sqrt{2}\pi \frac{N}{n}$. Thus N is some constant times n. For simplicity, we will assume $N = n$ for the remainder of the problem.
We will show that both terms are on the order $O(\frac{1}{\sqrt{n}})$.
For the first term:

$$\begin{aligned} \sum_{j=1}^{n-1} \theta_j \frac{\sqrt{2}\pi j}{n} &\leq \frac{\sqrt{2}\pi}{n} \sum_{j=1}^{n-1} \theta_j j \\ &\leq \frac{\sqrt{2}\pi}{n} \sqrt{\sum_{j=1}^{n-1} \theta_j^2 j^2} \sqrt{\sum_{j=1}^{n-1} 1} \\ &= O(\frac{1}{\sqrt{n}}) \end{aligned}$$

Where the second inequality follows from Cauchy-Schwartz.

For the second term:

$$\begin{aligned}\sum_{j=n}^{\infty} 2\theta_j &\leq \sum_{j=n}^{\infty} 2\frac{\theta_j j}{j} \\ &\leq \sqrt{\sum_{j=n}^{\infty} 2\theta_j^2 j^2}\sqrt{\sum_{j=n}^{\infty}\frac{1}{j^2}} \\ &\leq \frac{C}{\sqrt{n}}\end{aligned}$$

Where we use Cauchy-Schwartz and an integral approximation.

(b)

$$Var(\hat{\theta}_j) = \frac{1}{n^2}\sum_{i=1}^{n}\psi_j^2(x_i)Var(Y_i) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\psi_j^2(x_i)$$

Since the variance of Y_i is simply the variance of ϵ_i, because $m(x)$ is deterministic (i.e. $Y_i = m(x_i) + \epsilon_i$ and $Var(m(x_i)) = 0$). Then, each term in $\sum_{i=1}^{n}\psi_j^2(x_i)$ can be upper bounded by 2, since $0 \leq cos^2(\pi ji/n) \leq 1$, which results in:

$$Var(\hat{\theta}_j) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\psi_j^2(x_i) \leq \frac{2\sigma^2}{n}$$

This inequality is tight (in general), since for any j that is a multiple of $2n$, the sum is exactly $2n$. In other words, you cannot show that $Var(\hat{\theta}_j) = \frac{\sigma^2}{n}$ for all j, since if I take $j = 2n$ I get that $Var(\hat{\theta}_j) = \frac{2\sigma^2}{n}$. Of course, for asymptotics we probably don't care about the constant so it doesn't really make a difference.

(c) This quantity is just the risk under L_2 loss. We know that the risk decomposes as bias squared plus variance:

$$\mathbb{E}(\int_0^1 (\hat{m}(x) - m(x))^2 dx) = \int_0^1 b(\hat{m}(x))^2 + Var(m(x))$$

We know that:

$$\begin{aligned}b(\hat{m}(x)) &= \mathbb{E}\left[\sum_{j=0}^{J}\hat{\theta}_j\psi_j(x)\right] - \sum_{j=0}^{\infty}\theta_j\psi_j(x) = \sum_{j=0}^{J}\mathbb{E}\left[\hat{\theta}_j\right]\psi_j(x) - \sum_{j=0}^{\infty}\theta_j\psi_j(x) \\ &= -\sum_{j=J+1}^{\infty}\theta_j\psi_j(x)\end{aligned}$$

When we look at $b^2(\hat{m}(x))$, we have to multiply all of the terms in the sum, but we are actually interested in the integral:

$$\int_0^1 b^2(\hat{m}(x))dx = \sum_{k,j=J+1}^{\infty}\int_0^1 \theta_j\theta_k\psi_j(x)\psi_k(x)dx = \sum_{j=J+1}^{\infty}\theta_j^2$$

Since all of the cross terms cancel out due to orthogonality and since $\int_0^1 \psi_j^2(x)dx = 1$ for all j.

For the variance:

$$\int_0^1 Var(\hat{m}(x))dx = \sum_{j=0}^{J} \left(\int_0^1 \psi_j^2(x)dx \right) Var(\hat{\theta}_j) \leq \frac{2J\sigma^2}{n}$$

Using the result from before (which was different and not an equality statement), we get a slightly different result, but we're upper bounding the risk which is what's important, and we only differ by a constant factor. Putting these two quantities together we get the result, off by a constant factor.

(d) We'll find an upper bound on $\frac{J\sigma^2}{n} + \sum_{j=J+1}^{\infty} \theta_j^2$ that is independent of θ, and since that upper bound holds for all θ, it holds for the supremum.

First, writing $J = n^{\frac{1}{2\beta+1}}$ in the expression containing σ^2 gives $\sigma^2 n^{\frac{-2\beta}{2\beta+1}}$. We'll then bound the other term as follows:

$$J^{2\beta} \sum_{j=J+1}^{\infty} \theta_j^2 \leq \sum_{j=J+1}^{\infty} \theta_j^2 j^{2\beta} \leq \sum_{j=0}^{\infty} \theta_j^2 j^{2\beta} \leq C$$

Which means that the second term is upper bounded by $CJ^{-2\beta} = Cn^{\frac{-2\beta}{2\beta+1}}$. Thus both terms are upper bounded by some constant times $n^{\frac{-2\beta}{2\beta+1}}$ and adding the two together proves the result.

2. (a) The weights at time $t+1$ are:

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$$

Recursively expanding this for each t, (i.e. expanding $D_t(i)$) we'll get:

$$\begin{aligned} D_{t+1}(i) &= \frac{D_0(i) \exp\left\{-y_i \sum_{j=1}^{t} \alpha_j h_j(x_i)\right\}}{\prod_{j=0}^{t} Z_j} \\ &= \frac{\exp\{-y_i H(x_i)\}}{n \prod_{j=1}^{t} Z_j} \end{aligned}$$

Where we used the fact that $D_0(i) = \frac{1}{n}$ for all i and $Z_0 = 1$ (i.e. the weights are already normalized). Next, we use the fact that $\sum_{i=1}^{n} D_{t+1}(i) = 1$, and we bring $\prod_{j=1}^{t} Z_j$ to the other side to obtain the result:

$$1 = \sum_{i=1}^{n} \frac{\exp\{-y_i H(x_i)\}}{n \prod_{j=1}^{t} Z_j}$$
$$\frac{1}{n} \sum_{i=1}^{n} \exp\{-y_i H(x_i)\} = \prod_{j=1}^{t} Z_j$$

.

(b) First, we notice that we can decompse Z_t as (we'll focus just on one time step t):

$$\begin{aligned} Z_t &= \sum_{i=1}^{m} D_t(i)\exp(-\alpha_t y_i h_t(x_i)) = \sum_{i:y_i=h_t(x_i)} D_t(i)\exp(-\alpha_t) + \sum_{i:y_i\neq h_t(x_i)} D_t(i)\exp(\alpha_t) \\ &= \exp(-\alpha_t)\sum_{i:y_i=h_t(x_i)} D_t(i) + \exp(\alpha_t)\sum_{i:y_i\neq h_t(x_i)} D_t(i) \\ &= (1-\epsilon_t)\exp(-\alpha_t) + \epsilon_t\exp(\alpha_t) \end{aligned}$$

Where here we noticed that ϵ_t (i.e. the weighted training error at time t) is exactly equal to the sum of the weights of the data points that h_t classifies incorrectly. We'll take the derivative of this expression and set it equal to 0 to find an expression for α_t:

$$\begin{aligned} \frac{\partial Z_t}{\partial \alpha_t} &= -(1-\epsilon_t)\exp(-\alpha_t) + \epsilon_t\exp(\alpha_t) = 0 \\ \exp(2\alpha_t) &= \frac{1-\epsilon_t}{\epsilon_t} \\ \alpha_t &= \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t} \end{aligned}$$

Thus this choice of α_t minimizes Z_t and gives us the tightest upper bound on the training error. This is exactly the choice of α_t used in Adaboost.

(c) We have a class of hyperplanes in $\mathbb{R}^T$. If we consider a point x, we will represent it as a vector $(h_1(x),\ldots,h_T(x))$ and then we are just describing a hyperplane with coefficients $\alpha_1,\ldots,\alpha_T$ and with intercept fixed at 0. It is well known that the VC dimension of this set of functions is T (i.e. the dimension of the space). Using Sauer's Lemma, we immediately get that this set has shattering number $s_n \leq \left(\frac{en}{T}\right)^T$. This means that there are at most this many possible ways to partition n data points.

(d) Each $h_1,\ldots,h_T \in \mathcal{H}$ and since we choose T functions, there are exactly $|\mathcal{H}|^T$ ways to pick them. Thus the shattering number for $\mathcal{H}$ is $\leq |\mathcal{H}|^T\left(\frac{en}{T}\right)^T$. Using the VC theorem:

$$\begin{aligned} \sup_{H\in\mathcal{H}} \hat{R}(H) - R(H) &\leq \sqrt{\frac{8}{n}\log\left(\frac{4s(\mathcal{H},2n)}{\delta}\right)} \\ &\leq \sqrt{\frac{8}{n}\left(T\log|\mathcal{H}| + T\log\frac{2en}{T} + \log\frac{1}{\delta}\right)} \end{aligned}$$

Which is asymptotically the correct rate.

3. (a) We know the posterior is a dirichlet process with mean:

$$\bar{F}_n(x) = \frac{n}{n+\alpha}F_n(x) + \frac{\alpha}{n+\alpha}F_0(x)$$

When we compute the bias, the $F_n(x)$ in the first term will become $F(x)$ (because the empirical CDF is unbiased) and the second term is not random. Thus the bias is:

$$\begin{aligned}\mathbb{E}(\bar{F}_n(x)) - F(x) &= \frac{n}{n+\alpha}F(x) - F(x) + \frac{\alpha}{n+\alpha}F_0(x) \\ &= \frac{\alpha}{n+\alpha}\left(F_0(x) - F(x)\right)\end{aligned}$$

Which is pretty interesting, because as a decrease α, I decrease the bias of the estimator, which makes a lot of sense. Also, if F_0 is close to the truth, then my estimator has lower bias, which is pretty cool.

When we compute the variance, we use the fact that the first term is not random, and that the second term is just a scaled version of the empirical CDF:

$$\begin{aligned}Var(\bar{F}_n(x)) &= \frac{n^2}{(n+\alpha)^2}Var(F_n(x)) + 0 \\ &= \frac{n^2}{(n+\alpha)^2}\frac{1}{n}Var(\mathbb{1}(X < x)) \\ &= \frac{n}{(n+\alpha)^2}F(x)(1-F(x))\end{aligned}$$

Here we used standard properties of the variance, and the fact that the indicator function just specifies a Bernoulli random variable, with $p = F(x)$, and we know that bernoulli's have variance $p(1-p)$. This gives us the variance of the posterior mean.

The MSE of the posterior mean is the bias squared plus the variance. The empirical CDF is unbiased so its MSE is just its variance. Thus we look at the following:

$$\frac{\alpha^2}{(n+\alpha)^2}(F_0(x) - F(x))^2 + \frac{n}{(n+\alpha)^2}F(x)(1-F(x)) < \frac{1}{n}F(x)(1-F(x))$$

Whenever this inequality is satisfied, we know that the posterior mean has lower risk than the empirical CDF.

(b) The empirical CDF is easy. For a fixed x, we can view the empirical CDF as a sequence of n bernoulli random variables, each with $p = P(X < x)$. Moreover $F(x)$ is the mean, so we can immediately apply Hoeffding's inequality:

$$\mathbb{P}(F_n(x) - F(x) \geq \epsilon) < \exp\{-2n\epsilon^2\}$$

For the posterior mean, we have to do something more interesting. The main problem is $\bar{F}_n(x)$ is unbiased, but we can add terms to both sides (inside the probability), to account for this:

$$\begin{aligned}&\mathbb{P}(\bar{F}_n(x) - F(x) > \epsilon) \\ =\ &\mathbb{P}\left(\frac{n}{n+\alpha}F_n(x) - \frac{n}{n+\alpha}F(x) > \epsilon + \frac{\alpha}{n+\alpha}(F(x) - F_0(x))\right) \\ =\ &\mathbb{P}\left(F_n(x) - F(x) > \frac{n+\alpha}{n}\epsilon + \frac{\alpha}{n}(F(x) - F_0(x))\right) \\ \leq\ &\exp\left\{-2n\left(\frac{n+\alpha}{n}\epsilon + \frac{\alpha}{n}(F(x) - F_0(x))\right)^2\right\}\end{aligned}$$

Which gives you some (not particularly illuminating) concentration result for the posterior. The thing that is annoying is that the concentration depends on the deviation between the truth and the prior, which means that we need to select the prior well to get a good result.