18.657 PS 2 SOLUTIONS

1. Problem 1

(1) Let $Y_1, \ldots, Y_n$ be ghost copies of $X_1, \ldots, X_n$. Then we have

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i[f(X_i)-\mathbb{E}[f(X)]]\right|\right] = \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i[f(X_i)-\mathbb{E}[f(Y_i)]]\right|\right]$$
$$\leq \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i[f(X_i)-f(Y_i)]\right|\right],$$

by Jensen's inequality,

$$= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}[f(X_i)-f(Y_i)]\right|\right],$$

as $\sigma_i[f(X_i) - f(Y_i)]$ and $f(X_i) - f(Y_i)$ have the same distribution,

$$= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}[(f(X_i)-\mathbb{E}[f(X_i)])-(f(Y_i)-\mathbb{E}[f(Y_i)])]\right|\right]$$
$$\leq 2\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}[f(X_i)-\mathbb{E}[f(X_i)]]\right|\right],$$

by the triangle inequality.

(2) The distribution of g_i is the same as that of $|g_i|\sigma_i$, so we can write

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}g_i f(X_i)\right] = \mathbb{E}_{X_i,\sigma_i}\left[\mathbb{E}_{g_i}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}|g_i|\sigma_i f(X_i) \mid X_i,\sigma_i\right]\right]$$
$$\geq \mathbb{E}_{X_i,\sigma_i}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\mathbb{E}[|g_i|]\sigma_i f(X_i)\right],$$

by Jensen's inequality,

$$= \sqrt{\frac{2}{\pi}}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i f(X_i)\right],$$

using the first absolute moment of the standard Gaussian.

Date: Fall 2015.

(3) We compute:

$$
\begin{aligned}
R_n(\mathcal{F}+h) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i(f(X_i)+h(X_i))\right|\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f(X_i)\right|\right] + \mathbb{E}\left[\frac{1}{n}\left|\sum_{i=1}^n h(X_i)\right|\right],
\end{aligned}
$$

by the triangle inequality,

$$
\begin{aligned}
&= R_n(\mathcal{F}) + \mathbb{E}\left[\frac{1}{n}\left|\sum_{i=1}^n h(X_i)\right|\right], \\
&\le R_n(\mathcal{F}) + \frac{1}{n}\sqrt{\mathbb{E}\left[\left(\sum_{i=1}^n \sigma_i h(X_i)\right)^2\right]},
\end{aligned}
$$

by Jensen's inequality,

$$
\begin{aligned}
&= R_n(\mathcal{F}) + \frac{1}{n}\sqrt{\sum_{i=1}^n \mathbb{E}[h(X_i)^2] + 2\sum_{i<j}\mathbb{E}[\sigma_i\sigma_j h(X_i)h(X_j)]} \\
&= R_n(\mathcal{F}) + \frac{1}{n}\sqrt{\sum_{i=1}^n \mathbb{E}[h(X_i)^2]},
\end{aligned}
$$

by symmetry,

$$
\begin{aligned}
&\le R_n(\mathcal{F}) + \frac{1}{n}\sqrt{n\|h\|_\infty^2} \\
&= R_n(\mathcal{F}) + \frac{\|h\|_\infty}{n}.
\end{aligned}
$$

(4) By the triangle inequality, we have

$$
\begin{aligned}
R_n(\mathcal{F}_1+\ldots+\mathcal{F}_k) &= \mathbb{E}\left[\sup_{f_j\in\mathcal{F}_j} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i \sum_{j=1}^k f_j(X_i)\right|\right] \\
&\le \sum_{j=1}^k \mathbb{E}\left[\sup_{f_j\in\mathcal{F}_j} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i f_j(X_i)\right|\right] \\
&= \sum_{j=1}^k R_n(\mathcal{F}_j).
\end{aligned}
$$

(5) The supremum over different choices of f_j is at least the supremum over a single repeated choice:

$$\begin{aligned} R_n(\underbrace{\mathcal{F} + \ldots + \mathcal{F}}_{k}) &= \mathbb{E}\left[\sup_{f_j \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{k} f_j(X_i) \right|\right] \\ &\geq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \sum_{j=1}^{k} f(X_i) \right|\right] \\ &= k\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i f(X_i) \right|\right] \\ &= kR_n(\mathcal{F}), \end{aligned}$$

which provides the reverse bound to that of (4), establishing equality in this case.

2. Problem 2

(1) Let A be a maximum 2ε-packing of (T,d); let B be a minimum ε-covering. As B is an ε-covering, for each $a \in A$ there exists some choice of $b(a) \in B$ such that $d(a,b(a)) \le \varepsilon$. This map $b : A \to B$ is an injection: by the triangle inequality, no point in B can be within ε of two points of A, as these are at least 2ε apart. Thus $|A| \le |B|$, so that $D(T,d,2\varepsilon) \le N(T,d,\varepsilon)$.

Let C be maximum ε-packing. Note that C is in fact an ε-covering: if some point of T had distance greater than ε to each point of C, then we could add it to C to produce a larger ε-packing, contradicting maximality. It follows that $N(T,d,\varepsilon) \le D(T,d,\varepsilon)$.

(2) Let B^n denote the Euclidean ball in $\mathbb{R}^n$. Let $u \in B^n$ and $v \in B^m$. Then we can write $u = u_* + \varepsilon_u$, $v = v_* + \varepsilon_v$, where $u_* \in N_n$, $v_* \in N_m$, $\varepsilon_u \in \frac{1}{4}B^n$, and $\varepsilon_v \in \frac{1}{4}B^m$. We compute:

$$\begin{aligned}
\|M\| &= \sup_{u \in B^n, v \in B^m} u^\top M v \\
&= \sup_{u_*, v_*, \varepsilon_u, \varepsilon_v} u_*^\top M v_* + \varepsilon_u^\top M v_* + u^\top M \varepsilon_v \\
&\le \left(\sup_{u_* \in N_n, v_* \in N_m} u_*^\top M v_* \right) + \left(\sup_{\varepsilon_u \in \frac{1}{4}B^n, v_* \in N_m} \varepsilon_u^\top M v_* \right) + \left(\sup_{u \in B^n, \varepsilon_v \in \frac{1}{4}B^m} u^\top M \varepsilon_v \right) \\
&\le \left(\max_{u_* \in N_n, v_* \in N_m} u_*^\top M v_* \right) + \frac{1}{4}\|M\| + \frac{1}{4}\|M\|.
\end{aligned}$$

Rearranging, we have

$$\|M\| \le 2 \max_{u_* \in N_n, v_* \in N_m} u_*^\top M v_*$$

as desired.

(3) By independence and Hoeffding's lemma, we have, for all $s > 0$,

$$\begin{aligned}
\mathbb{E}[\exp(s\, u^\top M v)] &= \prod_{i,j} \mathbb{E}[\exp(s\, u_i M_{ij} v_j)] \\
&\le \prod_{i,j} \mathbb{E}\left[\exp\left(\frac{1}{2} s^2\, u_i^2 v_j^2 \right) \right] \\
&= \mathbb{E}\left[\exp\left(\frac{1}{2} s^2 \sum_{i,j} u_i^2 v_j^2 \right) \right] \\
&\le \exp(s^2/2),
\end{aligned}$$

whenever $u \in B^n$ and $v \in B^m$.

Recall from the notes that B^d has covering number at most $\left(\frac{3}{\varepsilon}\right)^d$, so that we are maximizing over $|N_n \times N_m| \le 12^{n+m}$ points. It follows from the standard maximal inequality for subgaussian random variables, or else by explicitly using Jensen with log and exp and replacing a maximum by a sum, that

$$\mathbb{E}\|M\| \le 2\sqrt{2\log(12^{m+n})} \le C(\sqrt{m} + \sqrt{n}).$$

3. Problem 3

(1) Let A be an ε-net for $[0,1]$; we can construct one with size at most $\frac{1}{2\varepsilon}+1$. Let X_{pre} denote the set of non-decreasing functions from $\{x_1,\ldots,x_n,\}$ to A, and let X be the set of functions $[0,1]\to[0,1]$ defined by piecewise linear extension of functions in X_{pre} (and constant extension on $[0,x_1]$ and $[x_n,1]$). It is straightforward to see that X is an ε-net for $(\mathcal{F},d_\infty^x)$: given $f\in\mathcal{F}$, we define $g\in X_{\mathrm{pre}}$ by taking $g(x_i)$ to be the least point of A lying within ε of $f(x_i)$, and then the function in X defined by extension of g is within ε of f.

It remains to count X. Let $a_1<\ldots<a_k$ be the elements of A, where $k\le\frac{1}{2\varepsilon}+1$. Functions $f\in X$ are uniquely defined by the count for each $1\le i\le k$ of how many x_j satisfy $f(x_j)=a_i$. A naive count of the possibilities for these values yields

$$N(\mathcal{F},d_\infty^x,\varepsilon)\le|X|\le(n+1)^{1+1/2\varepsilon}\le n^{2\varepsilon},$$

valid for $n\ge 3$.

(2) As $N(\mathcal{F},d_2^x,\varepsilon)\le N(\mathcal{F},d_\infty^x,\varepsilon)$, and $|f|\le 1$ for all $f\in\mathcal{F}$, the chaining bound yields

$$\begin{aligned}
\mathcal{R}_n &\le \inf_{\varepsilon>0} 4\varepsilon+\frac{12}{\sqrt{n}}\int_\varepsilon^1\sqrt{\log N(\mathcal{F},d_2^x,t)}\,\mathrm{d}t\\
&\le \inf_{\varepsilon>0} 4\varepsilon+\frac{12}{\sqrt{n}}\int_\varepsilon^1\sqrt{t^{-1}\log n}\,\mathrm{d}t\\
&= \inf_{\varepsilon>0} 4\varepsilon+24\sqrt{\frac{\log n}{n}}(1-\sqrt{\varepsilon})\\
&\le \lim_{\varepsilon\to0} 4\varepsilon+24\sqrt{\frac{\log n}{n}}(1-\sqrt{\varepsilon})\\
&= 24\sqrt{\frac{\log n}{n}}.
\end{aligned}$$

(3) We bound $N(\mathcal{F},d_1^x,\varepsilon)\le N(\mathcal{F},d_\infty^x,\varepsilon)\le n^{2/\varepsilon}$. The theorem in Section 5.2.1 yields

$$\begin{aligned}
\mathcal{R}_n &\le \inf_\varepsilon \varepsilon+\sqrt{\frac{2\log(2N(\mathcal{F},d_1^x,\varepsilon))}{n}}\\
&\le \inf_\varepsilon \varepsilon+\sqrt{\frac{2\varepsilon^{-1}\log n+2\log 2}{n}}.
\end{aligned}$$

Setting the two terms equal, to optimize over ε (in asymptotics), we have

$$\varepsilon^{3/2}=\sqrt{\frac{2\log n+2\varepsilon\log 2}{n}}\approx\sqrt{\frac{2\log n}{n}},$$

for a bound of

$$\mathcal{R}_n\lesssim\left(\frac{\log n}{n}\right)^{1/3},$$

which is strictly weaker than the chaining bound.

4. Problem 4

(1) We adapt the proof from class to the case of a penalized, rather than constrained, norm.

Let $\bar{W}_n$ be the span of the functions $k(x_i, -)$. We can decompose any function g uniquely as $g = g_n + g^\perp$, where $g_n \in \bar{W}_n$ and $g^\perp \perp \bar{W}_n$. As in class, $g^\perp(x_i) = \langle g^\perp, k(x_i, -)\rangle = 0$, so that $g(x_i) = g_n(x_i)$.

Plugging in to the objective function:

$$\psi(\mathbf{Y} - \mathtt{g}) + \mu\|g\|_W^2 = \psi(\mathbf{Y} - \mathtt{g}_n) + \mu\|g_n\|_W^2 + \mu\|g^\perp\|_W^2.$$

For any fixed g_n, this is a constant plus $\mu\|g^\perp\|_W^2$, which is minimized uniquely at $g^\perp = 0$. Thus (unfixing g_n) any minimizer $\hat{f}$ must lie in $\bar{W}_n$, as desired.

(2) As $\hat{f} \in \bar{W}_n$, write $\hat{f} = \sum_{i=1}^n \theta_i k(x_i, -)$, so that $\hat{\mathtt{f}} = K\theta$ and $\|\hat{f}\|_W^2 = \theta^\top K\theta$. Then the first-order optimality conditions read

$$\begin{aligned}
0 &= \nabla\left(\phi(\mathbf{Y} - \hat{\mathtt{f}}) + \mu\|\hat{f}\|_W^2\right) \\
&= \nabla\left(\mathbf{Y}^\top\Sigma^{-1/2}\mathbf{Y} - 2\mathbf{Y}^\top\Sigma^{-1/2}K\theta + \theta^\top K\Sigma^{-1/2}K\theta + \mu\theta^\top K\theta\right) \\
&= -2K\Sigma^{-1/2}\mathbf{Y} + 2K\Sigma^{-1/2}K\theta + 2\mu K\theta.
\end{aligned}$$

Rearranging, and recalling that $K\theta = \hat{\mathtt{f}}$, we have

$$(K\Sigma^{-1/2} + \mu I_n)\hat{\mathtt{f}} = K\Sigma^{-1/2}\mathbf{Y},$$

as desired.

(3) As above, it suffices to consider f, $g \in \bar{W}_n$, so that we can write $\mathtt{f} = K\theta_f$, $\mathtt{g} = K\theta_g$. From the definition of $\hat{f}$, we have

$$\psi(\mathtt{f} + \xi - \hat{\mathtt{f}}) + \mu\|\hat{f}\|_W^2 \le \psi(\mathtt{f} + \xi - \mathtt{g}) + \mu\|g\|_W^2.$$

Separating out the contribution of ξ, we obtain

$$\psi(\mathtt{f} - \hat{\mathtt{f}}) + \phi(\xi) + \xi^\top\Sigma^{-1/2}(\mathtt{f} - \hat{\mathtt{f}}) + \mu\|\hat{f}\|_W^2 \le \psi(\mathtt{f} - \mathtt{g}) + \psi(\xi) + \xi^\top\Sigma^{-1/2}(\mathtt{f} - \mathtt{g}) + \mu\|g\|_W^2.$$

Rearranging,

$$\begin{aligned}
\psi(\mathtt{f} - \hat{\mathtt{f}}) &\le -\xi^\top\Sigma^{-1/2}(\mathtt{f} - \hat{\mathtt{f}}) - \mu\|\hat{f}\|_W^2 + \psi(\mathtt{f} - \mathtt{g}) + \xi^\top\Sigma^{-1/2}(\mathtt{f} - \mathtt{g}) + \mu\|g\|_W^2 \\
&= \psi(\mathtt{f} - \mathtt{g}) + 2\mu\|g\|_W^2 + \xi^\top\Sigma^{-1/2}(\hat{\mathtt{f}} - \mathtt{g}) - \mu\|\hat{f}\|_W^2 - \mu\|g\|_W^2 \\
&\le \psi(\mathtt{f} - \mathtt{g}) + 2\mu\|g\|_W^2 + \xi^\top\Sigma^{-1/2}(\hat{\mathtt{f}} - \mathtt{g}) - \frac{\mu}{2}\|\hat{f} - g\|_W^2,
\end{aligned}$$

by the inequality $\|a - b\|_W^2 \le 2\|a\|_W^2 + 2\|b\|_W^2$. Let $Z = \Sigma^{-1/2}\xi$, which is distributed as $\mathcal{N}(0, I)$. Continuing the line of algebra,

$$\begin{aligned}
\psi(\mathtt{f} - \hat{\mathtt{f}}) &\le \psi(\mathtt{f} - \mathtt{g}) + 2\mu\|g\|_W^2 + Z^\top(\hat{\mathtt{f}} - \mathtt{g}) - \frac{\mu}{2}\|\hat{f} - g\|_W^2 \\
&= \psi(\mathtt{f} - \mathtt{g}) + 2\mu\|g\|_W^2 + Z^\top K(\theta_{\hat{f}} - \theta_g) - \frac{\mu}{2}(\theta_{\hat{f}} - \theta_g)^\top K(\theta_{\hat{f}} - \theta_g) \\
&\le \psi(\mathtt{f} - \mathtt{g}) + 2\mu\|g\|_W^2 + \frac{1}{\mu}Z^\top KZ,
\end{aligned}$$

by the inequality

$$a^\top Pb \le \frac{\mu}{2}a^\top Pa + \frac{1}{\mu}b^\top Pb$$

for any PSD matrix P. Now since

$$Z^\top KZ = \sum_{i,j=1}^n Z_i\langle k(x_i, -), k(x_j, -)\rangle_W = \|\sum_{i=1}^n Z_i k(x_i, -)\|_W^2,$$

we conclude by taking the infimum over $g \in \bar{W}_n$, which equals the infimum over $g \in W$.

(4) This follows from the previous part together with the observation

$$\mathbb{E}\left\|\sum_{i=1}^{n} Z_i k(x_i, -)\right\|_W^2 = \mathbb{E}\left[\sum_{i,j} Z_i Z_j K_{ij}\right] = \sum_{i=1}^{n} \mathbb{E}[Z_i^2] K_{ii} = \mathrm{Tr}(K).$$

(5) It is sufficient to prove the bound for $f \in \bar{W}_n$, since only its evaluations at the design points matter. For the Gaussian kernel we have $k(x,x) = 1$, so that $\mathrm{Tr}(K) = n$. Applying the previous part, and taking the case $g = f$ as an upper bound on the minimizer, we have

$$\mathbb{E}[\psi(\mathtt{f} - \hat{\mathtt{f}})] \leq \psi(\mathtt{f} - \mathtt{f}) + 2\mu\|f\|_W^2 + \frac{n}{\mu} = 2\mu\|f\|_W^2 + \frac{n}{\mu}.$$

Taking $\mu = \|f\|_W \sqrt{n/2}$ gives the result.

18.657 Mathematics of Machine Learning
Fall 2015