Homework 12 - Solution

1. (a) The first equality on the P-risk holds becaue

$$\begin{aligned}\mathbb{E}_P[(\mathbf{y}-f(\mathbf{x}))^2] &= \mathbb{E}_P[(\langle\theta_0,\mathbf{x}\rangle - f(\mathbf{x})-\varepsilon)]^2\\ &= \mathbb{E}_P[(\langle\theta_0,\mathbf{x}\rangle - f(\mathbf{x}))^2] + \mathbb{E}[\varepsilon^2] + 2\mathbb{E}_P[\varepsilon(\langle\theta_0,\mathbf{x}\rangle - f(\mathbf{x}))]\\ &= \mathbb{E}_P[(\langle\theta_0,\mathbf{x}\rangle - f(\mathbf{x}))^2] + \sigma^2 + 2\mathbb{E}_P[\varepsilon]\mathbb{E}_P[(\langle\theta_0,\mathbf{x}\rangle - f(\mathbf{x}))]\\ &= \mathbb{E}_P[(\langle\theta_0,\mathbf{x}\rangle - f(\mathbf{x}))^2] + \sigma^2,\end{aligned}$$

where we use the independence of ε and $\mathbf{x}$.
To find the Bayes predictor, we have to find f that minimizes $\mathcal{R}_P(f)$. Introduce the function $g_x(z) = (\langle\theta_0, x\rangle - z)^2 + \lambda z^2$. One can write $\mathcal{R}_P(f)$ as

$$\mathbb{E}_P[g_{\mathbf{x}}(f(\mathbf{x}))] + \sigma^2.$$

To minimize this quantity, we choose $f(x)$ as the minimizer of g_x for every x. One can check that this minimizer is equal to $\langle x, \theta_0/(1+\lambda)\rangle$.

(b) We have

$$\begin{aligned}\mathcal{R}_P^\star &= \mathcal{R}_P(f_P^\star)\\ &= \mathbb{E}_P[(\langle\theta_0,\mathbf{x}\rangle - \frac{\langle\theta_0,\mathbf{x}\rangle}{1+\lambda})^2] + \lambda\frac{\mathbb{E}_P[\langle\theta_0,\mathbf{x}\rangle^2]}{(1+\lambda)^2} + \sigma^2\\ &= \mathbb{E}_P[\langle\theta_0,\mathbf{x}\rangle^2]\left(\left(1-\frac{1}{1+\lambda}\right)^2 + \frac{\lambda}{(1+\lambda)^2}\right) + \sigma^2\\ &= \mathbb{E}_P[\langle\theta_0,\mathbf{x}\rangle^2]\frac{\lambda^2+\lambda}{(1+\lambda)^2} + \sigma^2\\ &= \mathbb{E}_P[\langle\theta_0,\mathbf{x}\rangle^2]\frac{\lambda}{1+\lambda} + \sigma^2.\end{aligned}$$

To conclude, we compute

$$\begin{aligned}\mathbb{E}_P[\langle\theta_0,\mathbf{x}\rangle^2] &= \mathbb{E}_P[\theta_0^\top\mathbf{x}\mathbf{x}^\top\theta_0]\\ &= \theta_0^\top\mathbb{E}_P[\mathbf{x}\mathbf{x}^\top]\theta_0\\ &= \theta_0^\top\mathrm{Id}_d\theta_0 = \|\theta_0\|^2.\end{aligned}$$

(c) This directly follows from the equality $\mathbb{E}_P[\langle\theta,\mathbf{x}\rangle^2] = \|\theta\|^2$, that holds for every $\theta\in\mathbb{R}^d$. We apply this identity to θ and $\theta-\theta_0$.

2. Let $(\mathbf{x_1},\mathbf{y_1}),\ldots,(\mathbf{x_n},\mathbf{y_n})$ be a sample of n i.i.d. observations from distribution P.

(a) The Hessian of the function is equal to $2(\lambda+1)\mathrm{Id}_d$. The function is therefore α-strongly convex for $\alpha = 2(\lambda+1)$.

(b) One can also write $\mathcal{R}_P(f_\theta)$ as

$$\mathbb{E}_P[\langle\theta_0-\theta,\mathbf{x}\rangle^2] + \lambda\|\theta\|^2 + \sigma^2.$$

The gradient $\nabla\mathcal{R}_P(f_\theta)$ is equal to

$$2\mathbb{E}[\mathbf{x}\langle\theta-\theta_0,\mathbf{x}\rangle] + 2\lambda\theta.$$

One can write $\mathbf{y_i} = \langle\theta_0,\mathbf{x_i}\rangle + \varepsilon_i$. Therefore,

$$\begin{aligned}\mathbf{v_i} &= 2\mathbf{x_i}(\langle\mathbf{x_i},\theta\rangle - \mathbf{y_i}) + 2\lambda\theta\\ &= 2\mathbf{x_i}\langle\mathbf{x_i},\theta-\theta_0\rangle - 2\mathbf{x_i}\varepsilon_i + 2\lambda\theta.\end{aligned}$$

As $\mathbb{E}[\varepsilon_i] = 0$ and ε_i is independent from $\mathbf{x_i}$, the expectation of this quantity is $\nabla\mathcal{R}_P(f_\theta)$, that is $\mathbf{v_i}$ is an unbiased estimate of $\nabla\mathcal{R}_P(f_\theta)$.

(c) It holds that (using the inequality $\|a+b\|^2 \le 2\|a\|^2 + 2\|b\|^2$)

$$\begin{aligned}\mathbb{E}[\|\mathbf{v_i}\|^2] &\le 8\mathbb{E}[\|\mathbf{x_i}(\langle\mathbf{x_i},\theta\rangle - \mathbf{y_i})\|^2] + 8\lambda^2\|\theta\|^2\\ &= 8\mathbb{E}[\|\mathbf{x_i}\|^2(\langle\mathbf{x_i},\theta\rangle - \mathbf{y_i})^2] + 8\lambda^2\|\theta\|^2\\ &\le 8M^2\mathbb{E}[(\langle\mathbf{x_i},\theta\rangle - \mathbf{y_i})^2] + 8\lambda^2\|\theta\|^2\\ &= 8M^2\mathbb{E}[(\langle\mathbf{x_i},\theta-\theta_0\rangle - \varepsilon_i)^2] + 8\lambda^2\|\theta\|^2\\ &= 8M^2(\mathbb{E}[(\langle\mathbf{x_i},\theta-\theta_0\rangle)^2] + \mathbb{E}[\varepsilon_i^2]) + 8\lambda^2\|\theta\|^2,\end{aligned}$$

where we use the fact that ε_i is centered and independent from $\mathbf{x_i}$ at the last line. It holds that $\mathbb{E}[\varepsilon_i^2] = \sigma^2$ and that $\mathbb{E}[(\langle \mathbf{x_i}, \theta - \theta_0 \rangle)^2] = \|\theta - \theta_0\|^2 \leq 4R^2$ (because both θ and θ_0 are in $B(0; R)$). We obtain the final bound

$$\mathbb{E}[\|\mathbf{v_i}\|^2] \leq 8M^2(4R^2 + \sigma^2) + 8\lambda^2 R^2$$

One can therefore apply stochastic gradient descent with projection on $B(0; R)$ using the vectors $(\mathbf{v_i})$. Theorem 1.5 in the lecture notes can be applied with $\rho = 8M^2(4R^2 + \sigma^2) + 8\lambda^2 R^2$ and $\alpha = 2(\lambda+1)$. According to this theorem, after n steps of stochastic gradient descent, the output $\hat{\theta}$ will satisfy

$$\mathcal{R}_P(f_\theta) - \mathcal{R}_P^\star \leq A \frac{\log n}{n}.$$

where A depends on the constants M, R and λ. The time complexity of this method is in $O(dn)$: there are n steps, and computing a single $\mathbf{v_i}$ requires $O(d)$ operations.