Machine Learning
SS 2020
Exercise sheet 1

Solution by
Daniel Strenger, Lorenzo Minneci
01531211, 11939539

April 11, 2020

Exercise 1. (Probability Refresher)

Solution. We calculate $\mathbb{P}(T=1, D=0) = \mathbb{P}(D=0)\mathbb{P}(T=1|D=0) = 95\% \cdot 1\% = 0.95\%$,
$\mathbb{P}(T=0, D=0) = \mathbb{P}(D=0)\mathbb{P}(T=0|D=0) = \mathbb{P}(D=0)(1-\mathbb{P}(T=1|D=0)) = 95\% \cdot 99\% = 94.05\%$,
$\mathbb{P}(T=1, D=1) = \mathbb{P}(D=1)\mathbb{P}(T=1|D=1) = 4\% \cdot 20\% = 0.8\%$,
$\mathbb{P}(T=0, D=1) = \mathbb{P}(D=1)(1-\mathbb{P}(T=1|D=1)) = 4\% \cdot 80\% = 3.2\%$,
$\mathbb{P}(T=1, D=2) = \mathbb{P}(D=2)\mathbb{P}(T=1|D=2) = 1\% \cdot 98\% = 0.98\%$,
$\mathbb{P}(T=0, D=2) = \mathbb{P}(D=2)(1-\mathbb{P}(T=1|D=2)) = 1\% \cdot 2\% = 0.02\%$,
$\mathbb{P}(T=1) = \mathbb{P}(T=1, D=0) + \mathbb{P}(T=1, D=1) + \mathbb{P}(T=1, D=2) = 2.73\%$ (because the events $D=0,1,2$ are disjoint),
$\mathbb{P}(T=0) = 1 - p(T=1) = 97.27\%$.
This gives the table

	healthy ($D=0$)	allergy ($D=1$)	celiac ($D=2$)	$p(T_i)$
pos. ($T=1$)	0.95%	0.8%	0.98%	2.73%
neg. ($T=0$)	94.05%	3.2%	0.02%	97.27%
$p(D_i)$	95%	4%	1%	

Exercise 2. (Mutual Information)

Solution. 1. We have

$$\Sigma^{-1} = \frac{1}{\det(\Sigma)} \begin{pmatrix} \sigma^2 & -\alpha\sigma^2 \\ -\alpha\sigma^2 & \sigma^2 \end{pmatrix} = \frac{1}{\sigma^4 - \alpha^2\sigma^4} \begin{pmatrix} \sigma^2 & -\alpha\sigma^2 \\ -\alpha\sigma^2 & \sigma^2 \end{pmatrix}$$

$$\Rightarrow (x,y)\Sigma^{-1} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{\sigma^4 - \alpha^2\sigma^4}(x,y) \begin{pmatrix} \sigma^2 & -\alpha\sigma^2 \\ -\alpha\sigma^2 & \sigma^2 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \frac{1}{\sigma^4(1-\alpha^2)}(x^2 + y^2 - 2xy\alpha)$$

$$\Rightarrow p_{X,Y}(x,y) = \frac{1}{2\pi\sigma^2\sqrt{1-\alpha^2}} \exp\left(-\frac{1}{2\sigma^2(1-\alpha^2)}(x^2 + y^2 - 2xy\alpha)\right)$$

The one dimensional marginal distributions of a multivariate normal distribution are univariate normal distributions,

$$p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right), \; p_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right).$$

From that we get

$$I(X;Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} p_{X,Y}(x,y) \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}\right) dydx$$

$$= \int_{\mathbb{R}} \int_{\mathbb{R}} p_{X,Y}(x,y) \log\left(\frac{\frac{1}{2\pi\sigma^2\sqrt{1-\alpha^2}} \exp\left(-\frac{1}{2\sigma^2(1-\alpha^2)}(x^2 + y^2 - 2xy\alpha)\right)}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{x^2}{2\sigma^2}\right) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)}\right) dydx$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} p_{X,Y}(x,y)\left(\log\left(\frac{1}{\sqrt{1-\alpha^2}}\right) - \frac{x^2+y^2-2xy\alpha}{2\sigma^2(1-\alpha^2)} + \frac{x^2}{2\sigma^2} + \frac{y^2}{2\sigma^2}\right)dydx$$

$$= -\frac{1}{2}\log(1-\alpha^2)\underbrace{\int_{\mathbb{R}}\int_{\mathbb{R}} p_{X,Y}(x,y)dydx}_{=1}$$

$$-\int_R\int_R p_{X,Y}(x,y)\left(\frac{x^2+y^2-2xy\alpha}{2\sigma^2(1-\alpha^2)} - \frac{x^2(1-\alpha^2)}{2\sigma^2(1-\alpha^2)} - \frac{y^2(1-\alpha^2)}{2\sigma^2(1-\alpha^2)}\right)dydx$$

$$= -\frac{1}{2}\log(1-\alpha^2) - \frac{\alpha}{1-\alpha^2}\int_{\mathbb{R}}\int_{\mathbb{R}} p_{X,Y}(x,y)\frac{\alpha x^2+\alpha y^2-2xy}{2\sigma^2}dydx$$

$$= -\frac{1}{2}\log(1-\alpha^2) - \frac{\alpha}{1-\alpha^2}\left(\int_{\mathbb{R}}\alpha(x-0)^2\underbrace{\int_{\mathbb{R}} p_{X,Y}(x,y)dy}_{=p_X(x)}dx + \int_{\mathbb{R}}\alpha(y-0)^2\underbrace{\int_{\mathbb{R}} p_{X,Y}(x,y)dx}_{=p_Y(y)}dy\right.$$

$$\left. -2\int_{\mathbb{R}}\int_{\mathbb{R}}(x-0)(y-0)p_{X,Y}(x,y)dydx\right)$$

$$= -\frac{1}{2}\log(1-\alpha^2) - \frac{\alpha}{1-\alpha^2}\left(\alpha\int_{\mathbb{R}}(x-\mu_x)^2p_X(x)dx + \alpha\int_{\mathbb{R}}(y-\mu_y)^2p_Y(y)dy - 2\int_{\mathbb{R}}\int_{\mathbb{R}}(x-\mu_x)(y-\mu_y)p_{X,Y}(x,y)dydx\right)$$

$$= -\frac{1}{2}\log(1-\alpha^2) - \frac{\alpha}{1-\alpha^2}\left(\alpha\mathbb{V}\mathrm{ar}(X) + \alpha\mathbb{V}\mathrm{ar}(Y) - 2\mathbb{C}\mathrm{ov}(X,Y)\right) = -\frac{1}{2}\log(1-\alpha^2) - \frac{\alpha}{1-\alpha^2}\left(\alpha\sigma^2 + \alpha\sigma^2 - 2\alpha\sigma^2\right)$$

$$= -\frac{1}{2}\log(1-\alpha^2)$$

2. $I(X;Y)$ is strictly monotonously increasing in α^2, so it is minimal for $\alpha^2 = 0 \Leftrightarrow \alpha = 0$, which corrsponds to the case that X and Y are independent. In this case $I(X;Y) = -\frac{1}{2}\log(1) = 0$, which also reflects the independece.
The value of $I(X;Y)$ increases as $\alpha^2 \to 1 \Leftrightarrow \alpha \to \pm 1$. As $|\alpha| < 1$ these values are never attained, but for $\alpha \to \pm 1$ the mutual information increases to infinity, so it is never maximal. The values $\alpha = \pm 1$ would correspond to the (degenerate) cases that Y is a function in X, i.e., if the value of X is given, then also Y is given.

3. The plots for $\sigma^2 \in \{1, 25, 200\}$ and $\alpha \in \{0, 0.5, 0.9\}$ are shown in figures 1-3.

4. As described previously the marginal densities do not depend on α. Also, because mean and variance for X and Y are equal, the plots for X and Y look the same for given σ^2. The plots are shown in figure 4.

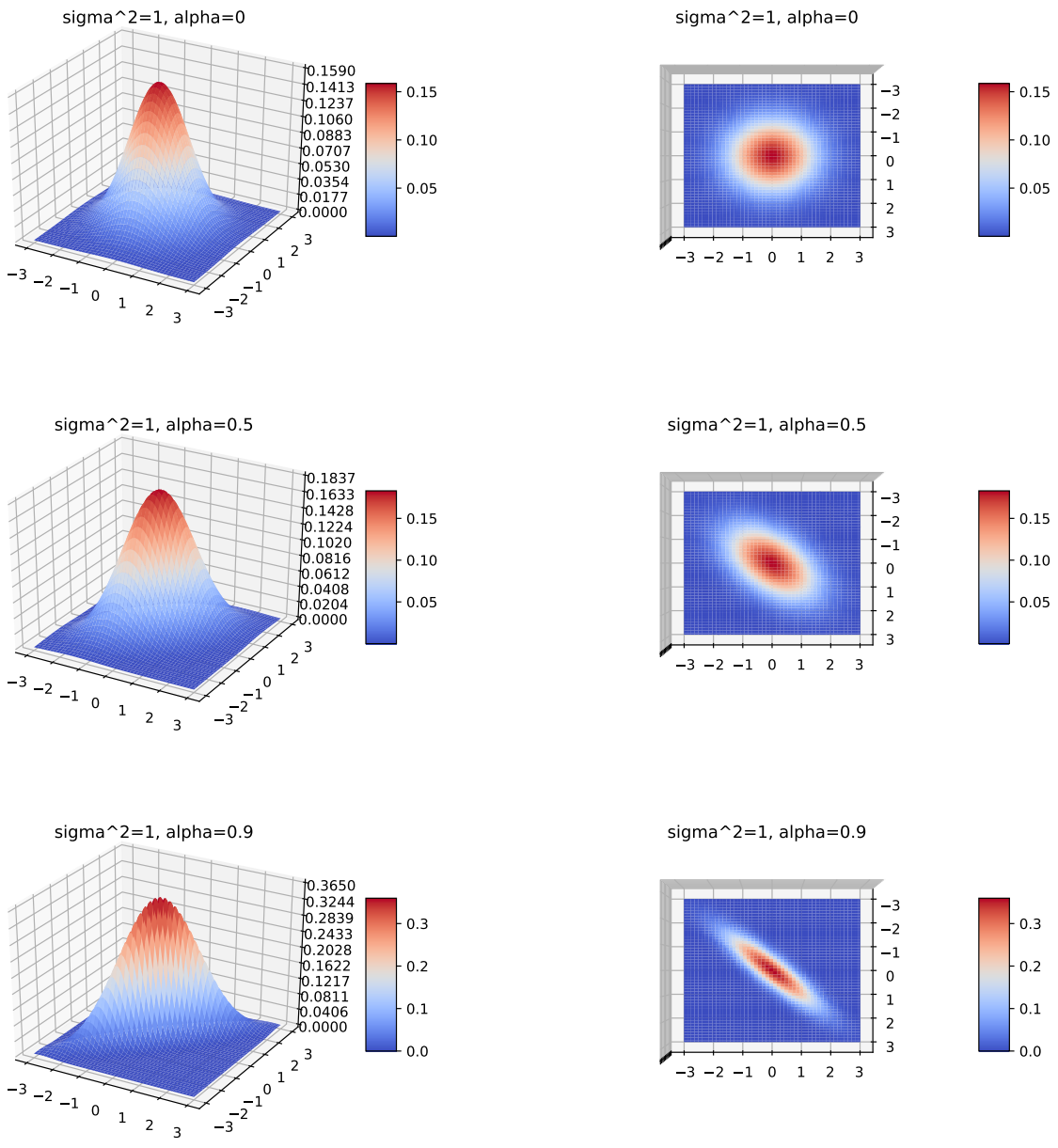

Figure 1: plots for $\sigma^2 = 1$

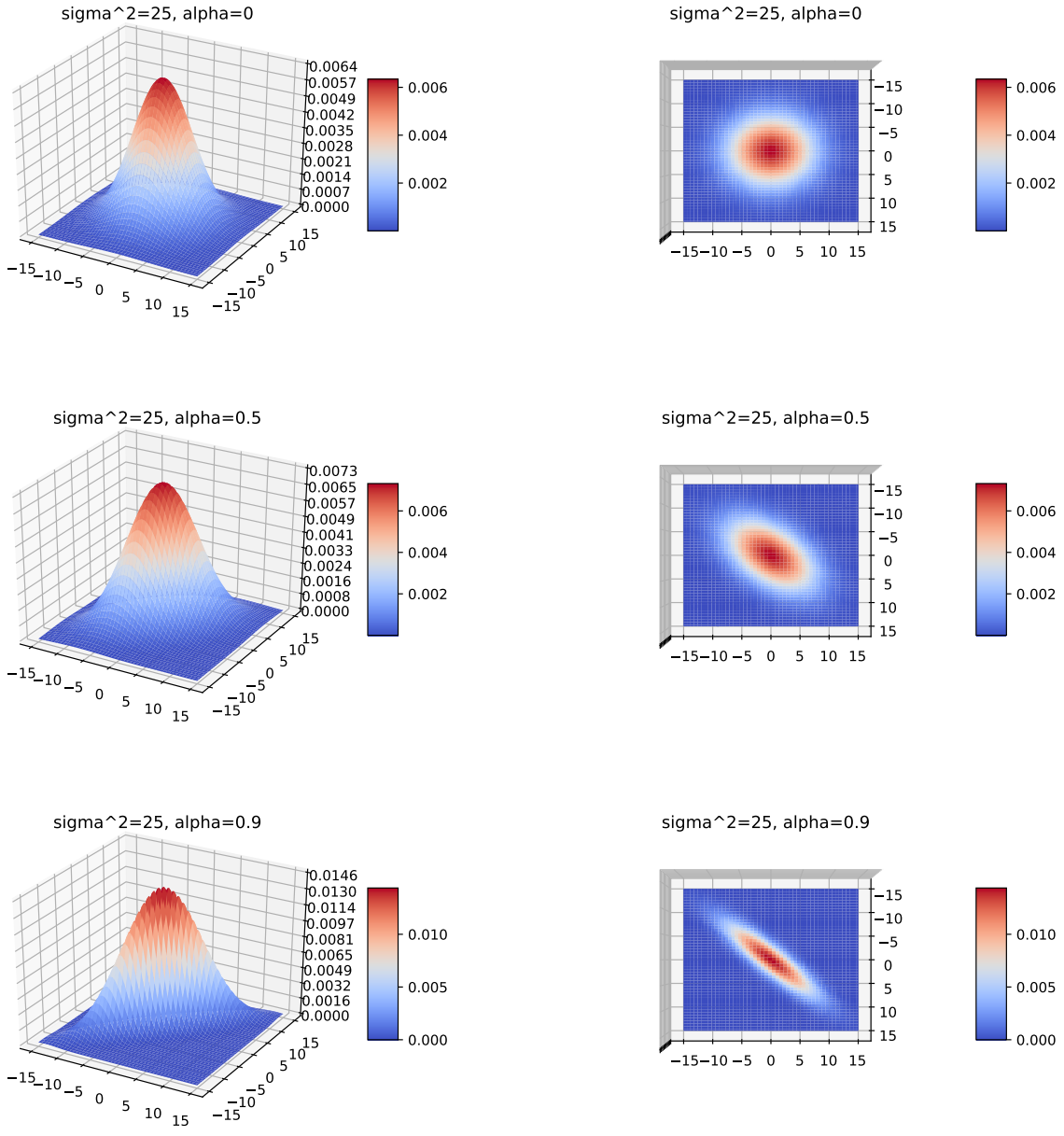

Figure 2: plots for $\sigma^2 = 25$

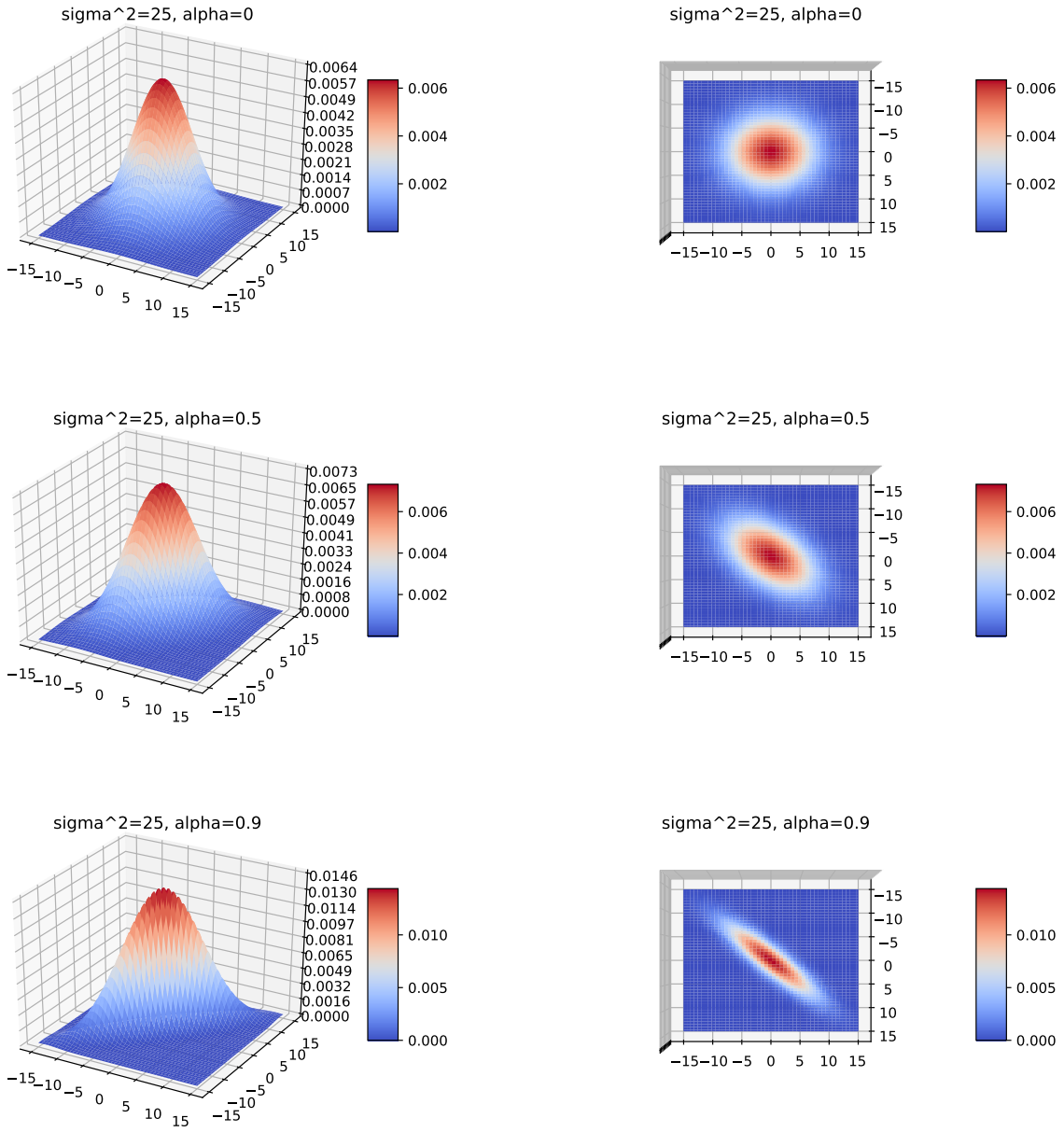

Figure 3: plots for $\sigma^2 = 100$

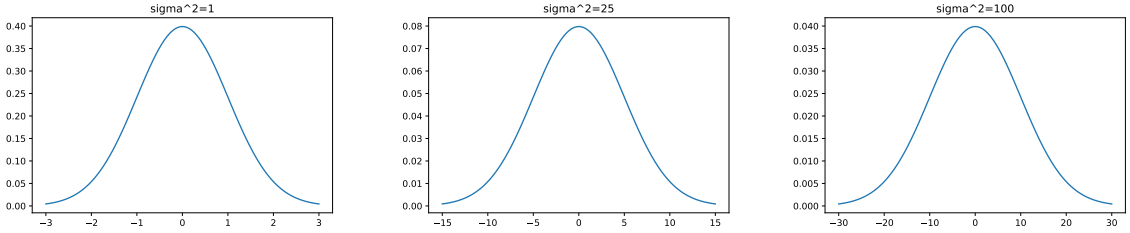

Figure 4: marginal densities of X and Y

Exercise 3. (Empirical Risk Minimization)

Solution. 1. We have

$$R(y) = \int\int (y(x)-t)^2 p_{X,T}(x,t)dtdx = \int\int (y(x) - \mathbb{E}(T|Xx + \mathbb{E}(T|x) - t)^2 p_{X,T}(x,t)dtdx$$

$$= \int\int (y(x)-\mathbb{E}(T|x))^2 p_{X,T}(x,t)dtdx + 2\int\int (y(x)-\mathbb{E}(T|x))(\mathbb{E}(T|x)-t)p_{X,T}(x,t)dtdx$$
$$+ \int\int (\mathbb{E}(T|x)-t)^2 p_{X,T}(x,t)dtdx$$

$$= \int\int (y(x)-\mathbb{E}(T|x))^2 p_{X,T}(x,t)dtdx + 2\int (y(x)-\mathbb{E}(T|x)) \underbrace{\int (\mathbb{E}(T|x)-t)\frac{p_{X,T}(x,t)}{p_X(x)}dt}_{=0} p_X(x)dx$$
$$+ \int \underbrace{\int (\mathbb{E}(T|x)-t)^2 \frac{p_{X,T}(x,t)}{p_X(x)}dt}_{=\mathbb{V}\mathrm{ar}(T|x)} p_X(x)dx.$$

$$= \int (y(x)-\mathbb{E}(T|x))^2 p_X(x)dx + 0 + \int \mathbb{V}\mathrm{ar}(T|x)p_X(x)dx \tag{1}$$

The choice of y has no impact on the second integral, so minimizing this expression is equivalent to minimizing the first integral. The integrand is non-negative, so the integral can be at least 0. This is the case iff $y(x) = \mathbb{E}(T|x)$ almost everywhere, so the choice $y(x) = \mathbb{E}(T|x)$ minimizes the expected loss. In our example we have

$$\mathbb{E}(T|x) = \int t p_T(t|x) = \int (\sin(x)+s)\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{s^2}{2\sigma^2})ds = \sin(x),$$

so $y^*(x) = \sin(x)$.

2. To find the minimizer w^*, consider

$$||\Phi w - t||^2 = (\Phi w - t)^T(\Phi w - t) = (\Phi w)^T(\Phi w) - (\Phi w)^T t - \underbrace{t^T(\Phi w)}_{=(\Phi w)^T t \text{ (scalar)}} + t^T t$$

$$= (\Phi w - t)^T(\Phi w - t) = w^T\Phi^T\Phi w - 2(\Phi w)^T t + t^T t$$

$$\Rightarrow \frac{\partial}{\partial w_i}||\Phi w - t||^2 = \frac{1}{N}\left(2 \underbrace{\Phi_{(i)}^T}_{\text{i-th line of } \Phi} \Phi w - 2\Phi_{(i)}^T t\right) \overset{!}{=} 0 \ \forall i = 1,...N$$

$$\Rightarrow \nabla_w||\Phi w - t||^2 = \frac{1}{N}\left(2\Phi^T\Phi w - 2\Phi^T t\right) \overset{!}{=} 0 \Leftrightarrow 2\Phi^T\Phi w = 2\Phi^T t \Leftrightarrow w = (\Phi^T\Phi)^{-1}\Phi^T t.$$

The matrix $\Phi^T\Phi$ is invertible, if all x_i are distinct, because in this case Φ has rank p, as its first p rows from a Vandermonde matrix. Moreover

$$\frac{\partial^2}{\partial w_i \partial w_j}||\Phi w - t||^2 = \frac{1}{N}2\Phi_{(i)}^T\Phi_{(j)},$$

which is positive definite, as Φ has full rank, so the found point $w^* = (\Phi^T\Phi)^{-1}\Phi^T t$is indeed a minimum.

3.-6. The plots for $N = 10$ are shown in figure 5, the empirical risks in figure 6. To compute the true risk, we use the decomposition of equation 1. We approximate the first integral by uniform sampling on $[0, 2\pi]$. To evaluate the second, note, that the conditional density $p_T(t|x)$ is the normal density of η (constant in x), so

$$\mathbb{V}\mathrm{ar}(T|x) = \int (t - \mathbb{E}(T|x))^2 p_T(t|x)dt = \int \eta^2 p_N(\eta)d\eta = \mathbb{V}\mathrm{ar}(\eta) = \sigma^2$$

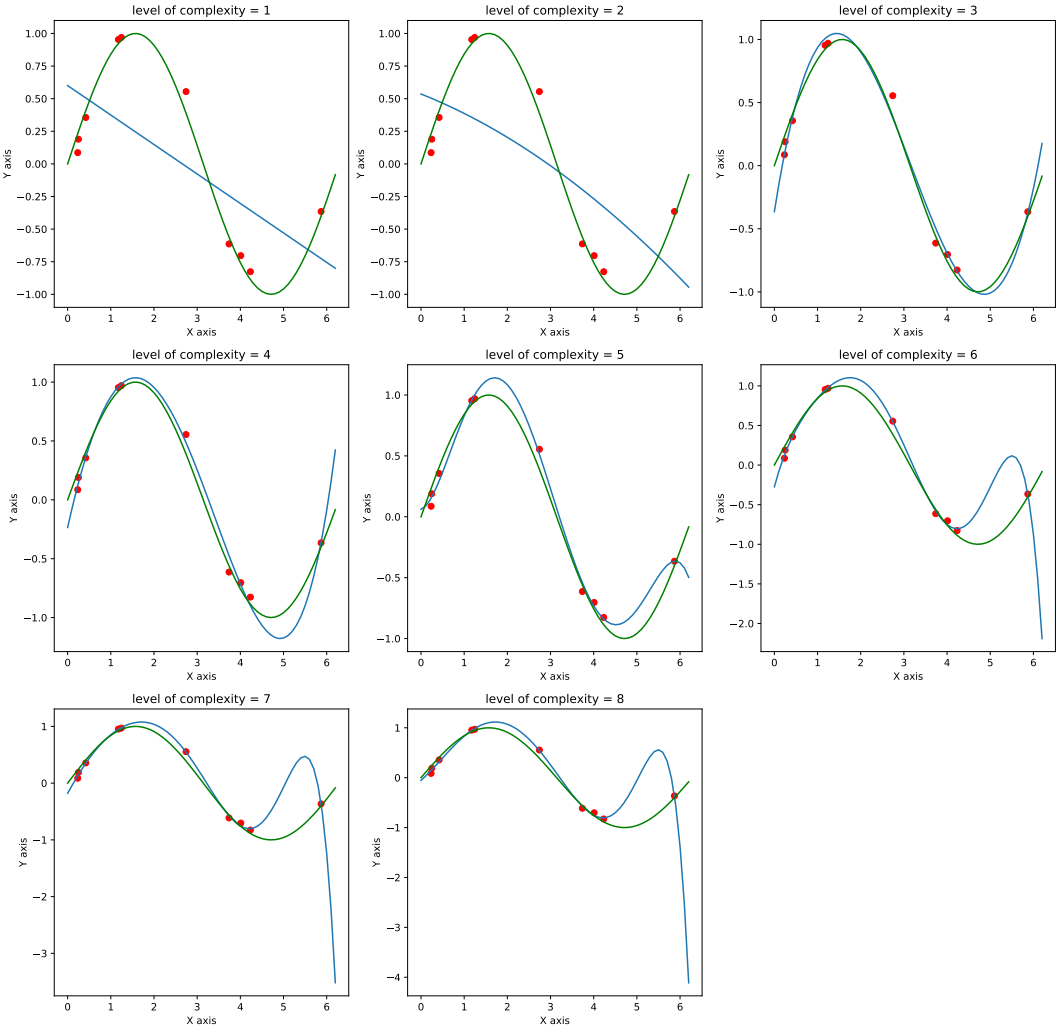

Figure 5: regression functions for $N = 10$ points

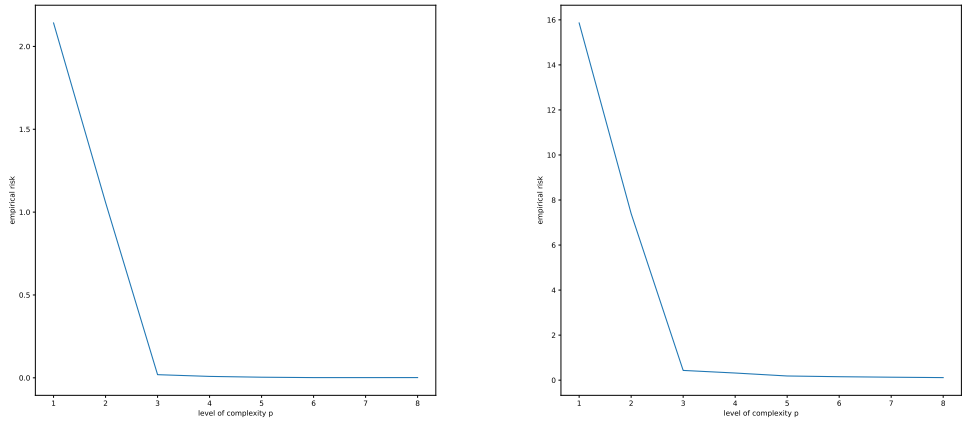

Figure 6: empirical risk for $N = 10$ and $N = 100$ points

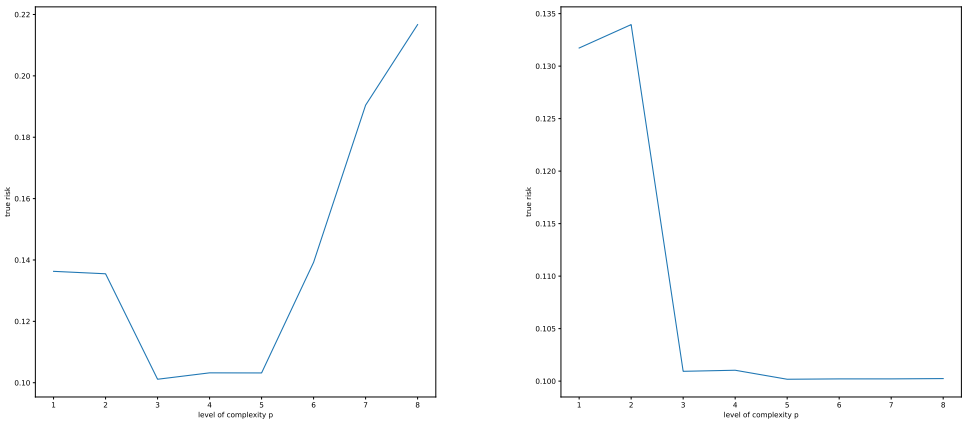

Figure 7: true risk for $N = 10$ and $N = 100$ points

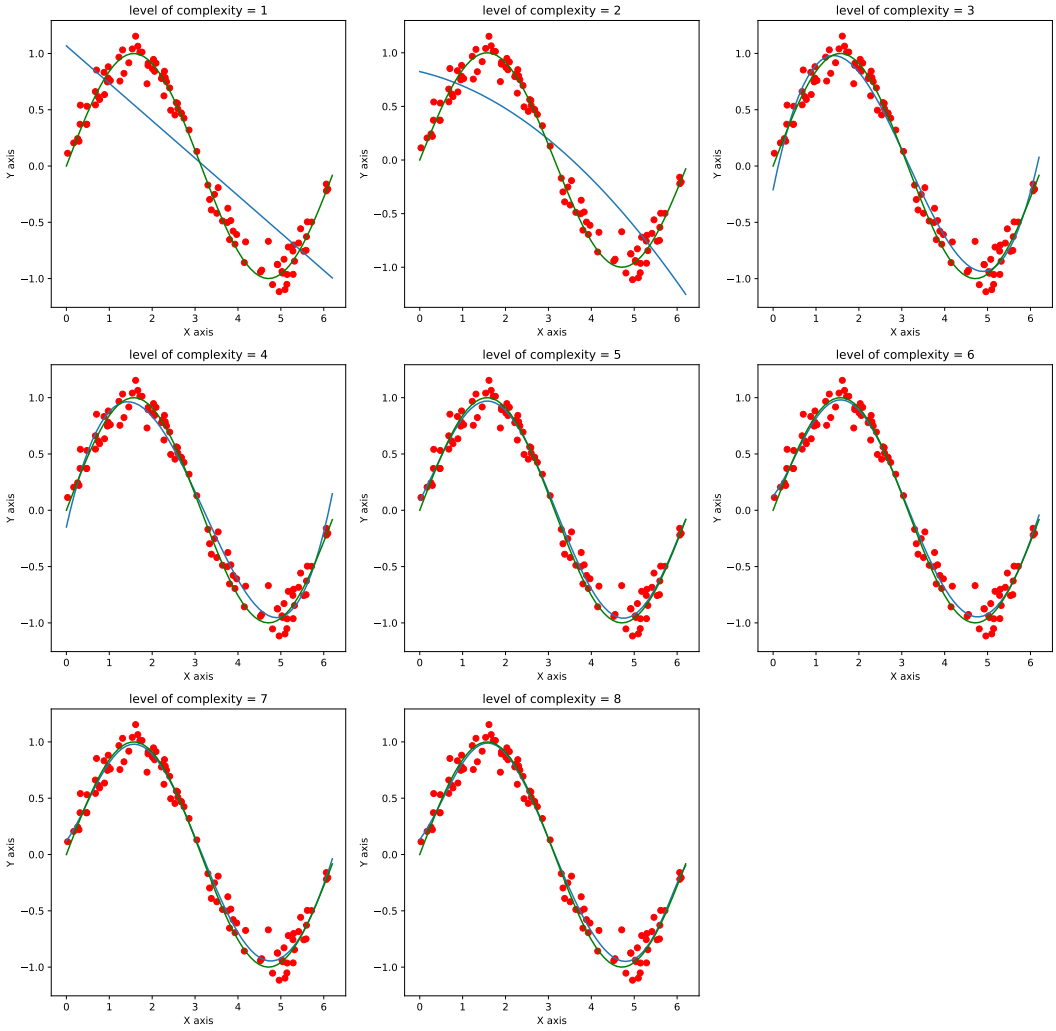

Figure 8: regression functions for $N = 100$ points

and therefore

$$\int \mathbb{V}\mathrm{ar}(T|x)p_X(x)dx = \sigma^2.$$

The plots of the true risks are given in figure 7. For both x-datasets the empirical risk decreases monotonously in the complexity level. This makes sense, because with a hight complexity level there are more degrees of freedom, so the polynomial can be chosen such that the curve is closer to the chosen points. For $N = 10$ the true risk is minimal for $p = 3$, for higher complexity the risk is higher again. The reason for this is overfitting. We have almost as many degrees of freedom as we have points to fit. The polynomial is calculated to fit the 10 chosen points as well as possible, but this comes at the cost of (almost) all other points on the curve. This is also visible in the plots, the regression function is very close to the red points, but for points, that lie a bit further away from the sample points, the curve isn't close at all. The quality of the regression here depends much on how the 10 sample points are distributed in the interval $[0, 2\pi]$. For different sample points (in different tests) the minimum was reached at a different complexity level, sometimes even at p=8.
For $N = 100$ the true risk decreases monotonously with a higher complexity level. In this case no overfitting occurs, because even for $p = 8$ there are many more points to be fitted than degrees of freedom. This is also visible in the regression plots for $N = 100$ (figure 8). The larger the complexity level is, the better the regression function fits the sample points.

Exercise 4. (MNIST - Bayesian Denoising)

Solution. We assume, that $p(Y = y_n) = p(Y = y_1) = c$ for all $n = 1, ..., N$, but no uniform distribution on $[0, 1]^D$, because that would mean $p(Y = y_n) = 0 \ \forall n = 1, ..., N$, which would lead to troubles in the definition of the conditional probability in the original sense.

1. We have

$$\frac{\sum_{n=1}^N y_n p(X = x|Y = y_n)}{\sum_{n=1}^N p(X = x|Y = y_n)} = \frac{\sum_{n=1}^N y_n p(X = x, Y = y_n)/p(Y = y_n)}{\sum_{n=1}^N p(X = x, Y = y_n)/p(Y = y_n)} = \frac{\frac{1}{c}\sum_{n=1}^N y_n p(X = x, Y = y_n)}{\frac{1}{c}\sum_{n=1}^N p(X = x, Y = y_n)}$$

$$= \frac{\sum_{n=1}^N y_n p(X = x, Y = y_n)}{\sum_{n=1}^N p(X = x, Y = y_n)} = \frac{\sum_{n=1}^N y_n p(X = x, Y = y_n)}{p(X = x)}$$

$$= \sum_{n=1}^N y_n p(Y = y_n|X = x) = \mathbb{E}_Y(Y|X = x) = \hat{y}_{CM}(x)$$

2. We have (because $\frac{c}{p(X=x)} > 0$)

$$\arg\max_{y_n} p(X = x|Y = y_n) = \arg\max_{y_n} \frac{c}{p(X = x)} p(X = x|Y = y_n) = \arg\max_{y_n} \frac{c}{p(X = x)} \frac{p(X = x, Y = y_n)}{c}$$

$$= \arg\max_{y_n} \frac{p(X = x, Y = y_n)}{p(X = x)} = \hat{y}_{MAP}(x)$$

3.-11. The original images are shown in figure 9, figures 10-12 show the noisy images and the ones denoised with conditional mean and denoised with MAP (in this order).

As one would expect, a higher variance results in more difficult to read noisy images. This also increases the number of wrongly denoised images.

The images denoised with with the CM tend to be more blury then the ones denoised with MAP (especially for strong noise). This comes from the fact, that MAP just selects an already existing image, so the results of MAP can only be as blurry as the blurriest given image. The results of CM are weighted superpositions of the images in Y_{test}, this leads to the not so clear images. This has the disadvantage that the results themselves might be difficult to read, but the fact that noisy images, for which it is very unclear, from which original number they come, result in images that look similar to more than one image in Y_{test} could also be seen as an advantage, depending on the application, because the uncertainty is visible in the result. If the MAP denoising selects one image it is impossible too see, whether the decision was very clear or there was a second image in Y_{test} (possibly reflecting a different number) that was almost equally likely as the chosen one. The results from the CM can be made less blurry by iterating, i.e., calculating the conditional mean w.r.t. the conditional mean of the original image and so on.

The CM approach seems to do more mistakes for high variance than the MAP approach.

Figure 9: original images

Figure 10: σ=0.25

Figure 11: σ=0.5

Figure 12: σ=1