Machine Learning B
Final Exam

101

January 2022

Contents

1 Perceptron

I know the main idea behind stochastic gradient descend is based on taking steps towards the negative gradient. This means that finding the gradient of the loss function is an obvious place to start. The hinge-loss has two "parts". For $yf(x) > 1$ the gradient is 0 by definition. For $yf(x) < 1$:

$$\frac{d}{d\omega} L_{hinge}(y, f(x)) = \frac{d}{d\omega} L_{hinge}(y, \langle \omega, x \rangle) = \frac{d}{d\omega}(1 - y\langle \omega, x \rangle) = -yx \tag{1}$$

So for every loop going towards negative gradient with a stepsize of η gives:

$$\begin{aligned} &\text{if } yf(x) < 1 \text{ then} \\ &\quad \omega \leftarrow \omega + \eta yx \\ &\text{else} \\ &\quad \omega \leftarrow \omega + \eta 0 \end{aligned}$$

As the result also corresponds to the same decision function, it is clear that this is the exact same thing as training a margin perceptron for $\eta = 1$:[1]

```
for every (x,y):
    if y<w,x> < 1 then
        w = w + yx
```

Exactly what we wanted!

2 Invariance properties of the polynomial kernel

My immediate reaction was no - of course a simple rotation would not change anything. Proving this on the other hand is harder:

The Representer Theorem tells us, that the solution can be written on the following form:

$$f(x) = \sum_{i=1}^{N} \beta_i k(x_i, x) + b \tag{2}$$

Since the β's are purely dependend on the classes (which won't change under a rotation), I simply need to show the invariance of the kernel to be able to assume invariance of the solution

$$k(\boldsymbol{x}, \boldsymbol{y}) = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^d \tag{3}$$

[1] Almost, at least: We were told to ignore the stopping criterion and the way the data-points are chosen

For a rotational matrix $\mathbf{R}$, the kernel looks as follows (assuming the standard dot-product):

$$k(\boldsymbol{Rx}, \boldsymbol{Ry}) = (\langle \boldsymbol{Rx}, \boldsymbol{Ry} \rangle + c)^d = ((\boldsymbol{Rx})^T \boldsymbol{Ry} + c)^d = (\boldsymbol{x^T R^T R y} + c)^d \tag{4}$$

Now using the fact that $\boldsymbol{R^T R} = 1$:

$$k(\boldsymbol{Rx}, \boldsymbol{Ry}) = (\langle \boldsymbol{Rx}, \boldsymbol{Ry} \rangle + c)^d = (\boldsymbol{x^T y} + c)^d = (\boldsymbol{x^T y} + c)^d = (\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^d \tag{5}$$

Now, just using the decision rule $h(x) = sgn(f(x))$ will grant us the same classifier whether or not the data is transformed (as long as it is simply rotated, at least).

3 Sparse classifiers

3.1

As we have an infinite hypothesis space, I remember Vapnik-Cherenkov and start out by looking at **Theorem 3.16**:

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{d_{\mathrm{VC}}} + 1\right) / \delta\right)}{n}} \right) \leq \delta \tag{6}$$

And as we know the VC-dimension of linear separators in $\mathbb{R}^d$ is simply d + 1, we can find the result, that with probability at least $1 - \delta$:

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{d+1} + 1\right) / \delta\right)}{n}} \tag{7}$$

3.2

Using the union bound and $d_{VC} = d + 1 = 2$

$$\mathbb{P}\left(\exists h \in \bigcup_{i=1}^{d} \mathcal{H}_i : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^2 + 1\right) / \delta\right)}{n}} \right) \leq$$

$$\sum_{i=1}^{d} \mathbb{P}\left(\exists h \in \mathcal{H}_i : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^2 + 1\right) / \delta\right)}{n}} \right) \leq$$

$$\sum_{i=1}^{d} \delta = d \cdot \delta$$

Doing some restructuring

$$\mathbb{P}\left(\exists h \in \mathcal{H}_{sparse} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{d \cdot 8 \ln \left(2 \left((2n)^2 + 1\right) / \delta\right)}{n}} \right) \leq \delta \tag{8}$$

ie. with probability at least $1 - \delta$

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{d \cdot 8 \ln \left(2 \left((2n)^2 + 1\right) / \delta\right)}{n}} \tag{9}$$

3.3

$$\sqrt{\frac{8\ln\left(2\left((2n)^{d+1}+1\right)/\delta\right)}{n}} = \sqrt{\frac{d\cdot 8\ln\left(2\left((2n)^{2}+1\right)/\delta\right)}{n}}$$
$$\ln\left(2\left((2n)^{d+1}+1\right)\right) = d\cdot\ln\left(2\left((2n)^{2}+1\right)\right)$$
$$(2n)^{d+1}+1 = e^{d}+(2n)^{2}+1$$

After having tried inserting different number and plotting this, I am quite convinced that the single sparse classifier is worse except for the unlikely cases when the dimensionality and amount of data is low. Which makes sense!

4 Bayesian inference of discrete distributions

4.1

This is basic statistics. The joint probability of seeing n q's is simply q times q n times.

$$\prod_{i}^{N_k} q_k = q_k^{N_k} \tag{10}$$

By same logic:

$$p = q_1^{N_1}\cdot q_2^{N_2}\cdot q_3^{N_3} = \prod_{i=1}^{3} q_k^{N_k} \tag{11}$$

4.2

Where the integral over the q'-vector $\bar{q}'$ is over all the possible values the different q's can take.

$$f_{Q|X}(\bar{q}|\bar{x}) = \frac{f_\Theta(\theta)p_{X|\Theta}(x\mid\theta)}{\int f_\Theta(\theta')\,p_{X|\Theta}(x\mid\theta')\,d\theta'} = \frac{f_Q(\bar{q})\prod_{k=1}^{3} q_k^{N_k}}{\int_{\bar{q}'} f_Q(\bar{q}')\prod_{k=1}^{3} {q'_k}^{N_k}\mathrm{d}\bar{q}'} \tag{12}$$

Here, both the posterior and prior distributions are continuous and the joint probability is discrete (in $\bar{x}$).

4.3

We are now given the prior $MLB(y,\alpha_1,\alpha_2,\alpha_3)$, which in our case looks as follows:

$$f_Q(\bar{q},\bar{\alpha}) = \frac{\Gamma\left(\sum_{k=1}^{3}\alpha_k\right)}{\prod_{k=1}^{3}\Gamma(\alpha_k)}\prod_{k=1}^{3} q_k^{\alpha_k-1} \tag{13}$$

Inserting this in the posterior distribution from the last question:

$$f_{Q|X}(\bar{q}|\bar{x}) = \frac{f_Q(\bar{q},\bar{\alpha})\prod_{k=1}^3 q_k^{N_k}}{\int_{\bar{q}'} f_Q(\bar{q}',\bar{\alpha})\prod_{k=1}^3 {q'_k}^{N_k}\mathrm{d}\bar{q}'} = \frac{\frac{\Gamma\left(\sum_{k=1}^3\alpha_k\right)}{\prod_{k=1}^3\Gamma(\alpha_k)}\prod_{k=1}^3 q_k^{\alpha_k-1}\prod_{k=1}^3 q_k^{N_k}}{\int_{\bar{q}'}\frac{\Gamma\left(\sum_{k=1}^3\alpha_k\right)}{\prod_{k=1}^3\Gamma(\alpha_k)}\prod_{k=1}^3 {q'_k}^{\alpha_k-1}\prod_{k=1}^3 {q'_k}^{N_k}\mathrm{d}\bar{q}'} \tag{14}$$

As $\frac{\Gamma\left(\sum_{k=1}^3\alpha_k\right)}{\prod_{k=1}^3\Gamma(\alpha_k)}$ is simply a numerical factor that does not depend on $\bar{q}$, this can be taken outside the integral and the whole distribution can be simplified as follows, where the two products are also combined:

$$\frac{\frac{\Gamma\left(\sum_{k=1}^3\alpha_k\right)}{\prod_{k=1}^3\Gamma(\alpha_k)}\prod_{k=1}^3 q_k^{\alpha_k-1}\prod_{k=1}^3 q_k^{N_k}}{\frac{\Gamma\left(\sum_{k=1}^3\alpha_k\right)}{\prod_{k=1}^3\Gamma(\alpha_k)}\int_{\bar{q}'}\prod_{k=1}^3 {q'_k}^{\alpha_k-1}\prod_{k=1}^3 {q'_k}^{N_k}\mathrm{d}\bar{q}'} = \frac{\prod_{k=1}^3 q_k^{N_k+\alpha_k-1}}{\int_{\bar{q}'}\prod_{k=1}^3 {q'_k}^{N_k+\alpha_k-1}\mathrm{d}\bar{q}'} \tag{15}$$

Now, focusing on the denominator, I will try evaluating the integral:

$$\int_{\bar{q}'}\mathrm{d}\bar{q}'\prod_{k=1}^3 {q'_k}^{E_k} = \int_{q_1}\mathrm{d}q_1\int_{q_2}\mathrm{d}q_2\int_{q_3}\mathrm{d}q_3(q_1^{E_1}q_2^{E_2}q_3^{E_3}) \tag{16}$$

Where $E_k = N_k + \alpha_k - 1$ and my background in physics shines through in the notation.

It is clear that q_1, q_2, q_3 cannot run from 0-1 each, as they are constrained by the equation $\sum_k q_k = 1$. Utilizing the symmetries of the problem, it is clear that one of the q's are trivially determined by the other two, making the integral limits easy to determine:

$$\int_0^1\mathrm{d}q_1\int_0^{1-q_1}\mathrm{d}q_2\ \ q_1^{E_1}q_2^{E_2}(1-q_1-q_2)^{E_3} \tag{17}$$

At first I spent quite a lot of time doing the integral by use of geometrical intuition and symmetries. Finally I simply threw the inner integral into WolframAlpha, which gave the following:

$$\int_0^{1-q_1}\mathrm{d}q_2\ \ q_1^{E_1}q_2^{E_2}(1-q_1-q_2)^{E_3} = \frac{\Gamma(E_2+1)\Gamma(E_3+1)}{\Gamma(E_2+E_3+2)}q_1^{E_1}(1-q_1)^{E_2+E_3+1} \tag{18}$$

This looked so very promising, so WolframAlpha got the honor of also calculating the second integral and, already having read the next question, allowing me to simplify :

$$\int_{\bar{q}'}\prod_{k=1}^3 {q'_k}^{E_k}\mathrm{d}\bar{q}' = \frac{\Gamma(E_2+1)\Gamma(E_3+1)}{\Gamma(E_2+E_3+2)}\int_0^1\mathrm{d}q_1\ \ q_1^{E_1}(1-q_1)^{E_2+E_3+1} \tag{19}$$

$$= \frac{\Gamma(E_2+1)\Gamma(E_3+1)}{\Gamma(E_2+E_3+2)}\cdot\frac{\Gamma(E_1+1)\Gamma(E_2+E_3+2)}{\Gamma(E_1+E_2+E_3+3)} \tag{20}$$

$$= \frac{\Gamma(E_1+1)\Gamma(E_2+1)\Gamma(E_3+1)}{\Gamma(E_1+E_2+E_3+3)} \tag{21}$$

$$= \frac{\prod_k\Gamma(E_k+1)}{\Gamma(\sum_k(E_k+1))} \tag{22}$$

Finally inserting this back into the posterior distribution:

$$f_{Q|X}(\bar{q}|\bar{x}) = \frac{\prod_{k=1}^{3} q_k^{E_k}}{\int_{\bar{q}'} \prod_{k=1}^{3} {q'_k}^{E_k} \mathrm{d}\bar{q}'} = \frac{\prod_{k=1}^{3} q_k^{E_k}}{\frac{\prod_k \Gamma(E_k+1)}{\Gamma(\sum_k (E_k+1))}} = \frac{\Gamma(\sum_k (E_k+1))}{\prod_k \Gamma(E_k+1)} \prod_{k=1}^{3} q_k^{E_k} \tag{23}$$

If you have been following along this should look very familiar to you, as we can define $\beta = E_k + 1 = N_k + \alpha_k$ and rewrite as follows:

$$f_{Q|X}(\bar{q}|\bar{x}) = \frac{\Gamma(\sum_k (\beta_k))}{\prod_k \Gamma(\beta_k)} \prod_{k=1}^{3} q_k^{\beta_k - 1} \tag{24}$$

Which is the exact **MLB**-distribution for $\alpha_k \leftarrow \alpha_k + N_k$, which is quite beautiful.

4.4

The definition of a conjugate prior distribution is "If posterior and prior belong to the same family of distributions, they are then called conjugate distributions, and prior is called a conjugate prior for the likelihood function."[2]

As we have:

$$f_{Q|X}(\bar{q}|\bar{x}, \bar{\alpha} + \bar{N}) = f_Q(\bar{q}, \bar{\alpha}) \tag{25}$$

There can be no doubt about the posterior distribution belonging to the same family of distributions as **MLB**, as it is simply an **MLB**-distribution with different parameters.

5 PAC-Bayes with distribution-dependent priors

5.1

Throughout this whole question I am using the expectation value notation $\mathbb{E}_\rho = \mathbb{E}_{h\sim\rho}$.

Start out by the definition of KL

$$KL(\rho||\pi) = \sum \rho \ln \frac{\rho}{\pi} = \sum \rho \ln \frac{Z_\pi e^{-F}}{Z_\rho e^{-G}} = \sum \rho \ln \frac{Z_\pi}{Z_\rho} + \sum \rho \ln \frac{e^{-F}}{e^{-G}} = \sum \rho \ln \frac{Z_\pi}{Z_\rho} + \sum \rho (G - F) \tag{26}$$

I then take the n'th hint and turn the sum into an expectation value (for now, mostly to ease on notation)

$$KL(\rho||\pi) = \mathbb{E}_\rho[\ln \frac{Z_\pi}{Z_\rho}] + \mathbb{E}_\rho[G - F] \tag{27}$$

[2]Taken directly from slide 10 of the "BayesianInference-MLB"

I can also take the other hint and turn the Z's from definition and, for reasons that will become apparent in due time, flip the sign and fraction in the logarithm.

$$\mathbb{E}_\rho[\ln \frac{Z_\pi}{Z_\rho}] + \mathbb{E}_\rho[G - F] = \mathbb{E}_\rho[\ln \frac{\rho e^{-G}}{\pi e^{-F}}] + \mathbb{E}_\rho[G - F] = \mathbb{E}_\rho[-\ln \frac{\pi e^{-F}}{\rho e^{-G}}] + \mathbb{E}_\rho[G - F] \quad (28)$$

$$\ln \frac{Z_\pi}{Z_\rho} = \ln \frac{\rho e^{-G}}{\pi e^{-F}} \quad (29)$$

I can now use the final hint and apply Jensen's Inequality ($-\ln(\cdot)$ is convex)

$$KL(\rho||\pi) \leq -\ln \mathbb{E}_\rho[\frac{\pi e^{-F}}{\rho e^{-G}}] + \mathbb{E}_\rho[G - F] \quad (30)$$

Now, remembering $\mathbb{E}_\rho x = \sum \rho x$, I can rewrite the expectation value:

$$KL(\rho||\pi) \leq -\ln \mathbb{E}_\pi[\frac{e^{-F}}{e^{-G}}] + \mathbb{E}_\rho[G - F] \quad (31)$$

I can now utilize a second Jensen's Inequality to pull e outside of the expectation value, giving us the final answer:

$$KL(\rho||\pi) \leq \mathbb{E}_\rho[G - F] - \mathbb{E}_\pi[G - F] \quad (32)$$

5.2

This proof has been left as an exercise for the grader.

5.3

Starting from the answer of 5.2, I simply insert the PAC-Bayes-λ upper bound (equation 3.22 from the Y. Seldin Machine Learning notes) in place of $\mathbb{E}[L(h)]$ and simplify[3]:

[3]Written verbatim from my handwritten notes - I wish you the best of luck following along

$$
\begin{aligned}
\frac{1}{\beta}\mathrm{KL}(\rho\|\pi) &\leq \frac{\lambda}{2}\mathbb{E}_\rho[L(h)] + \frac{\lambda}{2}\mathbb{E}_\rho[\hat{L}(h,S)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n\lambda} + \frac{\ln\frac{n+1}{\delta}}{n\lambda} \\
&\leq \frac{\lambda}{2}\frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)n} + \frac{\lambda}{2}\mathbb{E}_\rho[\hat{L}(h,S)] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n\lambda} + \frac{\ln\frac{n+1}{\delta}}{n\lambda} \\
&= \left(\frac{1}{n\lambda} + \frac{1}{n(2-\lambda)}\right)\mathrm{KL}(\rho\|\pi) + \frac{\lambda}{2}\frac{(2-\lambda/2)}{(1-\lambda/2)}\mathbb{E}_\rho[\hat{L}(h,S)] + \left(\frac{1}{n(2-\lambda)} + \frac{2}{n\lambda}\right)\ln\frac{n+1}{\delta} \\
&= \left(\frac{2-\lambda+\lambda}{n\lambda(2-\lambda)}\right)\mathrm{KL}(\rho\|\pi) + \frac{\lambda}{2}\frac{(4-\lambda)}{2(1-\lambda/2)}\mathbb{E}_\rho[\hat{L}(h,S)] + \left(\frac{\lambda+2(2-\lambda)}{n\lambda(2-\lambda)}\right)\ln\frac{n+1}{\delta} \\
&= \left(\frac{2}{n\lambda(2-\lambda)}\right)\mathrm{KL}(\rho\|\pi) + \frac{\lambda}{2}\frac{(4-\lambda)}{(2-\lambda)}\mathbb{E}_\rho[\hat{L}(h,S)] + \left(\frac{4-\lambda}{n\lambda(2-\lambda)}\right)\ln\frac{n+1}{\delta} \\
&= \frac{2\mathrm{KL}(\rho\|\pi)}{n\lambda(2-\lambda)} + \frac{\lambda(4-\lambda)\mathbb{E}_\rho[\hat{L}(h,S)]}{2(2-\lambda)} + \frac{(4-\lambda)\ln\frac{2(n+1)}{\delta}}{n\lambda(2-\lambda)}
\end{aligned}
$$

I do not know what more to write. It was so much easier to identify what to do in this question - this was the literal first thing I tried. Onwards and upwards!

5.4

To prove this I am starting where we left off, rearranging some terms:

$$
\begin{aligned}
\frac{1}{\beta}\mathrm{KL}(\rho\|\pi) &\leq \frac{2\mathrm{KL}(\rho\|\pi)}{n\lambda(2-\lambda)} + \frac{\lambda(4-\lambda)\mathbb{E}_\rho[\hat{L}(h,S)]}{2(2-\lambda)} + \frac{(4-\lambda)\ln\frac{2(n+1)}{\delta}}{n\lambda(2-\lambda)} \\
\mathrm{KL}(\rho\|\pi) &\leq \beta\left(\frac{2\mathrm{KL}(\rho\|\pi)}{n\lambda(2-\lambda)} + \frac{\lambda(4-\lambda)\mathbb{E}_\rho[\hat{L}(h,S)]}{2(2-\lambda)} + \frac{(4-\lambda)\ln\frac{2(n+1)}{\delta}}{n\lambda(2-\lambda)}\right) \\
\mathrm{KL}(\rho\|\pi) - \beta\frac{2\mathrm{KL}(\rho\|\pi)}{n\lambda(2-\lambda)} &\leq \beta\left(\frac{\lambda(4-\lambda)\mathbb{E}_\rho[\hat{L}(h,S)]}{2(2-\lambda)} + \frac{(4-\lambda)\ln\frac{2(n+1)}{\delta}}{n\lambda(2-\lambda)}\right)
\end{aligned}
$$

I will now start out by focusing on the left hand side of the inequality:

$$
\mathrm{KL}(\rho\|\pi) - \beta\frac{2\mathrm{KL}(\rho\|\pi)}{n\lambda(2-\lambda)} = \left(\frac{n\lambda(2-\lambda)-2\beta}{n\lambda(2-\lambda)}\right)\mathrm{KL}(\rho\|\pi)
$$

I will now be dividing by the factor in front of the $\mathrm{KL}(\rho\|\pi)$, but before we do this we have to be certain it is neither negative nor 0. As the denominator is always positive ($0 < \lambda < 2$), I will first check when the numerator is positive:

$$
n\lambda(2-\lambda) - 2\beta = -n\lambda^2 + 2n\lambda - 2\beta
$$

This is a 'sad' parabola[4], meaning it is only positive between its two roots. These can be found using high school math:

$$\lambda_{\pm} = \frac{-2n \pm \sqrt{4n^2 - 4n(2\beta)}}{-2n} = 1 \mp \sqrt{1 - 2\beta/n} \tag{33}$$

Now I have to check when the numerator makes the fraction 0, which is clear that it happens for $\lambda = 1$, $\beta = n/2$. This has given us the new bounds $\beta \in \left(0, \frac{n}{2}\right)$ and $\left(1 - \sqrt{1 - \frac{2\beta}{n}}, 1 + \sqrt{1 - \frac{2\beta}{n}}\right)$. I can now continue with the derivation, diving by the denumerator (aka. the whole fraction):

$$\begin{aligned}
\left(\frac{n\lambda(2-\lambda) - 2\beta}{n\lambda(2-\lambda)}\right) \mathrm{KL}(\rho\|\pi) &\leq \beta \left(\frac{\lambda(4-\lambda)\mathbb{E}_\rho[\hat{L}(h,S)]}{2(2-\lambda)} + \frac{(4-\lambda)\ln\frac{2(n+1)}{\delta}}{n\lambda(2-\lambda)}\right) \\
\mathrm{KL}(\rho\|\pi) &\leq \frac{\beta n\lambda(2-\lambda)}{n\lambda(2-\lambda) - 2\beta} \left(\frac{\lambda(4-\lambda)\mathbb{E}_\rho[\hat{L}(h,S)]}{2(2-\lambda)} + \frac{(4-\lambda)\ln\frac{2(n+1)}{\delta}}{n\lambda(2-\lambda)}\right) \\
\mathrm{KL}(\rho\|\pi) &\leq \frac{n\lambda\beta(4-\lambda)}{n\lambda(2-\lambda) - 2\beta} \left(\frac{\lambda\mathbb{E}_\rho[\hat{L}(h,S)]}{2} + \frac{\ln\frac{n+1}{\delta}}{n\lambda}\right)
\end{aligned}$$

This concludes the derivation.

5.5

Once again, I reach for the PAC-Bayes-λ upper bound. I start out by solving for $\mathrm{KL}(\rho\|\pi)$:

$$\begin{aligned}
\mathbb{E}_\rho[L(h)] &\leq \frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)n} \\
\left(\mathbb{E}_\rho[L(h)] - \frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}}\right)\lambda\left(1-\frac{\lambda}{2}\right)n - \ln\frac{n+1}{\delta} &\leq \mathrm{KL}(\rho\|\pi) \\
\mathbb{E}_\rho[L(h)]\lambda\left(1-\frac{\lambda}{2}\right)n - \mathbb{E}_\rho[\hat{L}(h,S)]\lambda n - \ln\frac{n+1}{\delta} &\leq \mathrm{KL}(\rho\|\pi)
\end{aligned}$$

As inserting this would only make the left hand side of the result from 5.4 smaller, I do so:

$$\mathbb{E}_\rho[L(h)]\lambda\left(1-\frac{\lambda}{2}\right)n - \mathbb{E}_\rho[\hat{L}(h,S)]\lambda n - \ln\frac{n+1}{\delta} \leq A\left(\frac{\lambda\mathbb{E}_\rho[\hat{L}(h,S)]}{2} + \frac{\ln\frac{n+1}{\delta}}{n\lambda}\right) \tag{34}$$

[4]:(

Where I have optimized my LaTeX-clutter by introducing the shorthand $A = \frac{n\lambda\beta(4-\lambda)}{n\lambda(2-\lambda)-2\beta}$.

Knowing where I want to end up, I will now massage this inequality until it looks the way I want it to. I do this by first solving for $\mathbb{E}_\rho[L(h)]$:

$$\mathbb{E}_\rho[L(h)]\lambda\left(1-\frac{\lambda}{2}\right)n - \mathbb{E}_\rho[\hat{L}(h,S)]\lambda n - \ln\frac{n+1}{\delta} \leq A\left(\frac{\lambda\mathbb{E}_\rho[\hat{L}(h,S)]}{2} + \frac{\ln\frac{n+1}{\delta}}{n\lambda}\right)$$

$$\mathbb{E}_\rho[L(h)]\lambda\left(1-\frac{\lambda}{2}\right)n - \mathbb{E}_\rho[\hat{L}(h,S)]\lambda n \leq A\left(\frac{\lambda\mathbb{E}_\rho[\hat{L}(h,S)]}{2} + \frac{\ln\frac{n+1}{\delta}}{n\lambda}\right) + \ln\frac{n+1}{\delta}$$

$$\mathbb{E}_\rho[L(h)] \leq \frac{\left(A\left(\frac{\lambda\mathbb{E}_\rho[\hat{L}(h,S)]}{2} + \frac{\ln\frac{n+1}{\delta}}{n\lambda}\right) + \ln\frac{n+1}{\delta}\right) + \lambda n\mathbb{E}_\rho[\hat{L}(h,S)]}{\lambda\left(1-\frac{\lambda}{2}\right)n}$$

Now I combine the terms involving $\mathbb{E}_\rho[\hat{L}(h,S)]$ and $\ln\frac{n+1}{\delta}$ before reintroducing A and simplifying:

$$\mathbb{E}_\rho[L(h)] \leq \frac{\left(A\left(\frac{\lambda\mathbb{E}_\rho[\hat{L}(h,S)]}{2} + \frac{\ln\frac{n+1}{\delta}}{n\lambda}\right) + \ln\frac{n+1}{\delta}\right) + \lambda n\mathbb{E}_\rho[\hat{L}(h,S)]}{\lambda\left(1-\frac{\lambda}{2}\right)n}$$

$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho[\hat{L}(h)]}{1-\frac{\lambda}{2}}\left(1+\frac{A}{2\lambda n}\right) + \frac{\ln\frac{n+1}{\delta}}{n\lambda(1-\frac{\lambda}{2})}\left(1+\frac{A}{n\lambda}\right)$$

$$\mathbb{E}_\rho[L(h)] \leq \frac{\mathbb{E}_\rho[\hat{L}(h,S)]}{1-\frac{\lambda}{2}}\left(1+\frac{\beta\lambda(4-\lambda)}{2(n\lambda(2-\lambda)-2\beta)}\right) + \frac{\ln\frac{n+1}{\delta}}{n\lambda\left(1-\frac{\lambda}{2}\right)}\left(1+\frac{\beta(4-\lambda)}{n\lambda(2-\lambda)-2\beta}\right)$$

This is the exact answer we were looking for!

5.6 closing remarks

After inserting the definitions for $\rho(h)$ and $\pi(h)$ giving me expectation values over $G(h) - F(h,S)$ I have no idea what to do in 5.2. I have spent more time on that one sub-question than I have the rest of the exam combined - I spent literal days!

6 6 Confidence intervals for Bernoulli

6.1

The Hoeffding bound I use is the classic two-sided we all know and love. With a probability of $1-\delta$:

$$\left|\frac{1}{n}\sum_{i=1}^{n}X_i - \mu\right| \leq \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \tag{35}$$

This is easy to use, as I can simply insert the number of data points n and the chosen $\delta = 0.05$, and calculate both the upper and lower bound.

The Empirical Bernstein for Bernoulli variables are taken directly from Sadegh's "Concentration Inequalities" slides. Again, with probability $1 - \delta$:

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \sqrt{\frac{2V_n}{n} \log(4/\delta)} + \frac{7 \log(4/\delta)}{3(n-1)} \tag{36}$$

Again, we have a symmetrical bound. This time, a third variable is in use. The empirical variance V_n has to be calculated. Again turning to Sadegh's slides, I know this can be calculated as:

$$V_n = \frac{1}{n(n-1)} \sum_{1 \leq i < j \leq n} (X_i - X_j)^2 \tag{37}$$

I am unsure what to include in this report, so here is a double-for-loop snippet from my Notebook[5]:

```
def sample_variance(x):
    s = 0
    for i in range(len(x)):
        for j in range(i+1,len(x)):
            s += (x[i]-x[j])**2

    return s/(n*n-n)
```

Now, using the KL-divergence as bound on the means are slightly more annoying. But with some massaging, a numerical solution can be found as:

$$\text{Upper bound} = \sup \left\{ p : \text{kl}\left(\hat{p}_n \| p\right) \leq \frac{\ln(n+1)/\delta}{n} \right\}$$
$$\text{Upper bound} = \inf \left\{ p : \text{kl}\left(\hat{p}_n \| p\right) \leq \frac{\ln(n+1)/\delta}{n} \right\}$$

We did this exact thing in a hand-in, so the idea is taken from there. The $\frac{\ln(n+1)/\delta}{n}$ term is taken from equation 2.11 in Yevgenys lecture notes.

As we have seen this before, I simply stole my own code, the main component of which is a binary search. In semi-pseudo-code the following happens:

```
while True:
    p = (pmin + pmax)/2

    diff = kl(p_emp, p) - z
```

[5]Which can be found in the zip-file I handed in

```
    if diff == 0:
        return p
    elif diff > 0:
        pmax = p
    else:
        pmin = p
```

The bounds I ended up getting are:

S	Lower bound	Upper bound	Width
Hoeffding	0.703	0.876	0.173
Empirical Bernstein	0.671	0.908	0.237
KL-divergence	0.671	0.882	0.2112

Table 1: Lots of numbers

6.2

For comparing the three confidence intervals, I thought some visual aid would help. Here is a plot:

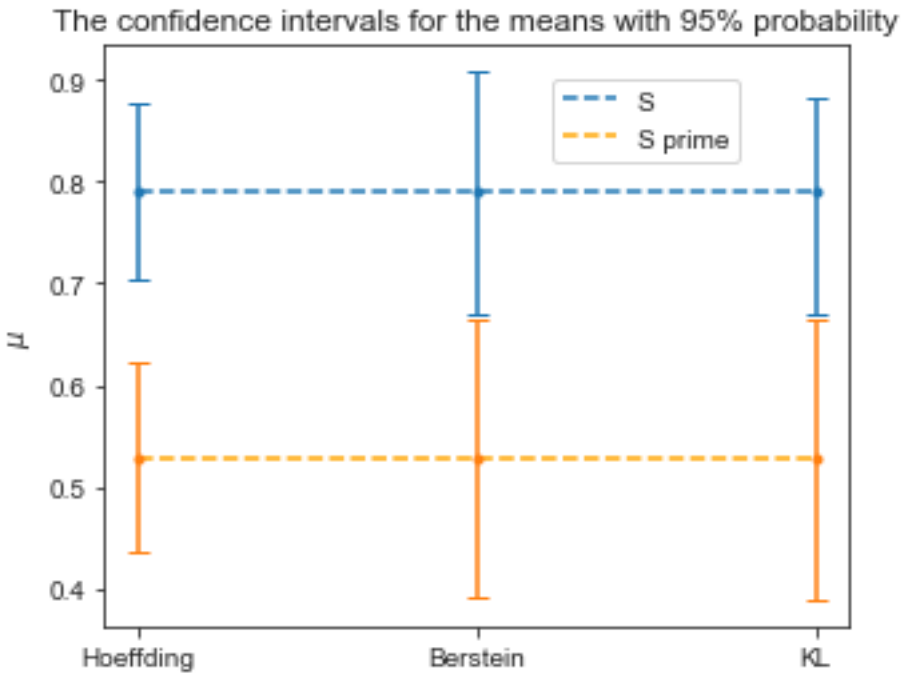

Figure 1: Yengev

As was apparent in Table 1, Hoeffding is the tightest bound. On a strong second place, the KL-inequality has a comparable upper bound, but the lower bound has dropped quite a lot. Lastly the Empirical Bernstein has the same lower bound as KL,

But as we do not know the analytical. the width of the is the most important measure. They agree with what I write KL OR empirical best for low mu

6.3

Again, Looking at a visualization, we can easily see:

S	Lower bound	Upper bound	Width
Hoeffding	0.686	0.893	0.207
Empirical Bernstein	0.643	0.936	0.294
KL-divergence	0.671	0.882	0.211

S'	Lower bound	Upper bound	Width
Hoeffding	0.416	0.641	0.225
Empirical Bernstein	0.360	0.697	0.337
KL-divergence	0.389	0.665	0.276

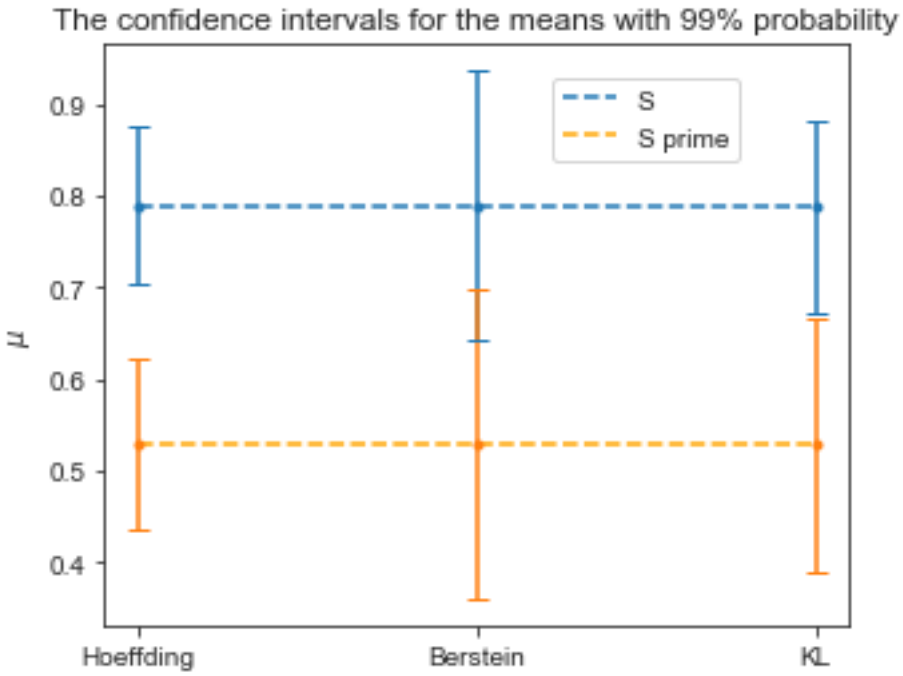

Figure 2: Yengev

Using Hoeffding or the KL-inequality, one can easily be convinced, that with a 99% probability the means are different.

Seeing the bounds as uncertainties, I choose to take the Hoeffding bound and add the uncertainties in quadrature as we are looking at a difference between two values:

$$\sigma_{\mu-\mu'} = \sqrt{\sigma_\mu^2 + \sigma_{\mu'}^2} = \sqrt{(0.207/2)^2 + (0.225/2)^2} = 0.153 \tag{38}$$

As the difference between the two means are:

$$\mu - \mu' = 0.789 - 0.529 = 0.26 \tag{39}$$

I would argue that with a probability of 99% μ is above μ', that is:

$$\mu - \mu' = 0.26 \pm 0.15 \tag{40}$$