IEMS 402 Statistical Learning - 2025 Winter
HW1

Tuorui Peng[1]

Exercise 1 Design of Loss Function

1.(a)

Consider decomposing $m(X)$ as follows:

$$m(X) := \mathbb{E}\left[Y|X=x\right] + \delta(X),$$

i.e. with $\delta(X)$ being the deviation of the true conditional expectation from the model. Then the expected ℓ_2 error can be written as:

$$\begin{aligned}
\mathbb{E}\left[(Y-m(X))^2\right] =& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right]-\delta(X))^2\right] \\
=& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])^2\right] + \mathbb{E}\left[\delta^2(X)\right] - 2\mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])\delta(X)\right] \\
\overset{(i)}{=}& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])^2\right] + \mathbb{E}\left[\delta^2(X)\right] \\
\overset{(ii)}{\geq}& \mathbb{E}\left[(Y-\mathbb{E}\left[Y|X=x\right])^2\right].
\end{aligned}$$

Thus proved that the expected ℓ_2 error is minimized by the conditional expectation $\mathbb{E}\left[Y|X=x\right]$. Here in the proof, (i) is due to the fact that $\mathbb{E}_{Y|X=x}\left[Y-\mathbb{E}\left[Y|X=x\right]\middle|X=x\right]=0$, and (ii) is due to the non-negativity of $\mathbb{E}\left[\delta^2(X)\right]$, and equality holds if and only if $\delta(X)=0$ almost surely.

1.(b)

The expected ℓ_1 error can be written as:

$$\mathbb{E}\left[|Y-m(X)|\right] = \mathbb{E}_X\left[\int_{Y|X=x} |Y-m(x)|\,\mathrm{d}F(y|X=x)\right]$$

taking variation with respect to $m(X)$, we have (here δ refers to the variation operator):

$$\begin{aligned}
\delta\mathbb{E}\left[|Y-m(X)|\right] =& \mathbb{E}_X\left[\int_{Y|X=x} \delta m(x)\cdot\operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x)\right] \\
=& \mathbb{E}_X\left[\delta m(x)\int_{Y|X=x} \operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x)\right]
\end{aligned}$$

To minimize the expected ℓ_1 error, we requre the variation taking value of zero, i.e.:

$$0=\delta\mathbb{E}\left[|Y-m(X)|\right] = \mathbb{E}_X\left[\delta m(x)\int_{Y|X=x} \operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x)\right],\quad \forall\delta m(\,\cdot\,)$$

which requires choosing $m(\,\cdot\,)$ s.t. $\int_{Y|X=x}\operatorname{sgn}(Y-m(x))\,\mathrm{d}F(y|X=x) = \mathbb{E}_{Y|X=x}\left[\operatorname{sgn}(Y-m(x))|X=x\right] = 0$ almost surely. This is equivalent to using the conditional median as the prediction function $m(x) = \operatorname{median}(Y|X=x)$.

[1] TuoruiPeng2028@u.northwestern.edu

1.(c)

Similarly we write the following differentiation w.r.t. β

$$\frac{\partial}{\partial\beta}\mathbb{E}\left[(Y-\beta' X)^2\right] = -\mathbb{E}\left[-2X(Y-\beta' X)\right]$$

and set it to zero, we have:

$$\begin{aligned} 0 =& -\mathbb{E}\left[-2X(Y-\beta' X)\right] \\ \Rightarrow \beta_* =& \mathbb{E}\left[XX'\right]^{-1}\mathbb{E}\left[XY\right] \end{aligned}$$

1.(d)

We consider the following function:

$$s_\alpha(y,\hat{y}) := \begin{cases} \alpha, & \text{if } y-\hat{y}>0 \\ \alpha-1, & \text{if } y-\hat{y}<0 \end{cases}$$

and notice that $s_\alpha(y,\hat{y}) = \dfrac{\partial}{\partial y}\alpha\cdot\operatorname{sgn}(y-\hat{y})$[2]. Then we can write the deviation of expected ℓ_α error as:

$$\delta\mathbb{E}\left[\rho_\alpha(y,m(x))\right] = \mathbb{E}_X\left[\delta m(x)\int_{Y|X=x} s_\alpha(y,m(x))\,\mathrm{d}F(y|X=x)\right]$$

minimizing the ρ_α loss function requires the variation to be zero for any δm, i.e.:

$$0 = \mathbb{E}_{Y|X=x}\left[s_\alpha(y,m(x))|X=x\right] \Rightarrow m(x) = q_\alpha(x)$$

where $q_\alpha(x)$ is the α-quantile of Y given $X=x$.

Exercise 2 Central Limit Theorem

2.(a)

First by SLLN we definitely have $\bar{X} \xrightarrow{\text{a.s.}} \mathbb{E}[X] = \mu$.

Then we re-write the expression of s_n^2 as:

$$\begin{aligned} s_n^2 =& \frac{1}{n}\sum_{i=1}^n (X_i-\bar{X})^2 \\ =& \frac{1}{n}\sum_{i=1}^n X_i^2 - \bar{X}^2 \end{aligned}$$

[2]Omitting the discontinuity at $y-\hat{y}=0$, which won't be a big problem if we have continuous and strictly increasing loss function.

Then we notice that:

$$\text{by LLN: } \frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{p} \mathbb{E}\left[X^2\right]$$

$$\text{by LLN and continuous mapping theorem: } \bar{X}^2 \xrightarrow{p} \left(\mathbb{E}\left[X\right]\right)^2$$

thus we have by slutsky's theorem:

$$s_n^2 \xrightarrow{\mathrm{d}} \mathbb{E}\left[X^2\right] - \left(\mathbb{E}\left[X\right]\right)^2 = \sigma^2$$

2.(b)

Note that $s_n^2 \xrightarrow{\mathrm{d}} \sigma^2$ which is a constant, thus we also have $s_n \xrightarrow{\mathrm{p}} \sigma$. Then by Slutsky's theorem and CLT, we have the following:

$$\begin{aligned}
&\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{\mathrm{d}} N(0,1)\\
&\frac{s_n}{\sigma} \xrightarrow{\mathrm{p}} 1\\
\Rightarrow &\frac{\sqrt{n}(\bar{X}_n - \mu)}{s_n} \xrightarrow{\mathrm{d}} N(0,1)
\end{aligned}$$

Exercise 3 Curse of Dimensionality: Asymptotic scaling of nearest neighbor distances

3.(a)

$$\begin{aligned}
\mathbb{P}\left(\left\|x_{i(X_0)} - x_0\right\| > \delta\right) =& \mathbb{P}\left(\bigcap_{i=1}^{n}\{\|x_i - x_0\|_2 > \delta\}\right)\\
=& \int \mathrm{d}P_{x_0} \int \mathrm{d}P_{x_1^n} \prod_{i=1}^{n} \mathbf{1}_{\|x_i - x_0\|_2 > \delta}\\
=& \int \left(1 - P\left(B_d(x,\delta)\right)\right)^n \mathrm{d}P(x)
\end{aligned}$$

3.(b)

We can construct the partition as follows: at each dimention, construct cutting points $\{-kr, (-k+1)r, \ldots, (k-1)r, kr\}$ where k chosen s.t. $k = \lceil \frac{R}{\delta/d} \rceil \leq 2Rd/\delta$ and $r = \delta/d$. And each $U_\cdot$ is constructed by the combination of the cutting points. Then we have number of partition

$$N(\delta) = (2k+1)^d \leq \frac{8(Rd)^d}{\delta^d} = \frac{c}{\delta^d}$$

in this way, each "block" of the partition is at most a hypercube with side length r, and diameter diam $= r\sqrt{d} < \delta$.

3.(c)

Since the partition $U_1^{N(\delta)}$ has diameter at most δ, we consider THE block $U_\imath$ that contains x for each given x. Then we have:

$$U_\imath \subseteq B_d(x,\delta) \Rightarrow P\left(U_\imath\right) \leq P\left(B_d(x,\delta)\right)$$

thus we have:

$$\begin{aligned}
\mathbb{P}\left(\left\|x_{i(X_0)} - x_0\right\| > \delta\right) &= \int \left(1 - P\left(B_d(x,\delta)\right)\right)^n \mathrm{d}P(x) \\
&\leq \sum_{i=1}^{N(\delta)} \int_{U_i} \left(1 - P\left(P\left(U_i\right)\right)\right)^n \mathrm{d}P(x) \\
&= \sum_{i=1}^{N(\delta)} \left(1 - P\left(P\left(U_i\right)\right)\right)^n P\left(U_i\right) \\
&\overset{(i)}{\leq} \frac{c}{en\delta^d}.
\end{aligned}$$

Thus finished the proof. Here in the proof, (i) is due to the fact that $x \mapsto x(1-x)^n$ reaches maximum at $x = 1/(n+1)$, with maximum value

$$\frac{1}{n+1}(1 - \frac{1}{n+1})^n = \frac{1}{n}\left(1 - \frac{1}{n+1}\right)^{n+1} \leq \frac{1}{en}.$$

3.(d)

With the probabilistic bound, we note that to maintain the bound at $O(1)$, we should choose $\delta \asymp n^{-1/d}$ (so that $c/en\delta^d = O(1)$). Which indicates that

$$\mathbb{P}\left(\left\|x_{i(X_0)} - x_0\right\| \lesssim n^{-1/d}\right) \geq 1 - C$$

i.e. with certain minimal probability, the nearest neighbor distance is at most $\lesssim n^{-1/d}$.

Exercise 4

4.(a)

Note that $f_\theta(x) = 0$ is a hyper plane in $\mathbb{R}^d$, the distance from $x^{(i)}$ to which is

$$\begin{aligned}
\text{distance} &= \frac{\left|\theta' x^{(i)} + \theta_0\right|}{\|\theta\|} \\
&= \begin{cases} \dfrac{\left|\theta' x^{(i)} + \theta_0\right|}{\|\theta\|}, & \text{if } \theta' x^{(i)} + \theta_0 > 0 \\ -\dfrac{\left|\theta' x^{(i)} + \theta_0\right|}{\|\theta\|}, & \text{if } \theta' x^{(i)} + \theta_0 < 0 \end{cases}
\end{aligned}$$

further for hard margin SVM, $\theta' x^{(i)} + \theta_0$ has the same sign as $y^{(i)}$, thus we have:

$$\text{distance} = \frac{y^{(i)}(\theta' x^{(i)} + \theta_0)}{\|\theta\|} = \gamma^(i)$$

4.(b)

Optimization problem for hard margin SVM can be written as:

$$\begin{aligned}&\arg\min_{\theta,\theta_0}\frac{1}{2}\|\theta\|^2\\&w.r.t.y^{(i)}(\theta' x^{(i)}+\theta_0)\geq 1\end{aligned}$$

and note that the decision boundary is determined by the (θ,θ_0), which has an extra degree of freedom w.r.t. scale transformation. We cancel this degree of freedom by setting a constraint $\|\theta\|=\frac{1}{M}$.

The Lagrangian can be written as:

$$\mathcal{L}(\theta,\theta_0,\alpha_1^n)=\frac{1}{2}\|\theta\|^2-\sum_{i=1}^{n}\alpha_i\left(y^{(i)}(\theta' x^{(i)}+\theta_0)-1\right)$$

and the optimization problem can be solved by minimizing the Lagrangian w.r.t. θ,θ_0

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\theta}=&\theta-\sum_{i=1}^{n}\alpha_i y^{(i)}x^{(i)}=0\\\frac{\partial\mathcal{L}}{\partial\theta_0}=&-\sum_{i=1}^{n}\alpha_i y^{(i)}=0\end{aligned}$$

to get the dual problem:

$$\begin{aligned}\theta_D(\alpha)=&-\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y^{(i)}y^{(j)}x^{(i)'}x^{(j)}+\sum_{i=1}^{n}\alpha_i\\&s.t.\alpha_i\geq 0,\sum_{i=1}^{n}\alpha_i y^{(i)}=0\end{aligned}$$

The dual problem is usually easier to solve because we can see that the dual problem has less (non-trivial) constraints.

IEMS 402 Statistical Learning - 2025 Winter
HW2

Tuorui Peng[1]

Exercise 1 Adaptive to Anisotropic Smoothnes

We write the $\hat{m}(x)$ as follows:

$$\begin{aligned}\hat{m}(x) =& \frac{\int y\hat{p}(x,y)\,\mathrm{d}y}{\hat{p}(x)}\\ =& \frac{\sum_{i=1}^n h^{-2}K(\frac{X_i-x}{h})\int yK(\frac{Y_i-y}{h})\,\mathrm{d}y}{\sum_{i=1}^n h^{-1}K(\frac{X_i-x}{h})}\\ =& \frac{\sum_{i=1}^n h^{-2}K(\frac{X_i-x}{h})\int (y-Y_i+Y_i)K(\frac{Y_i-y}{h})\,\mathrm{d}y}{\sum_{i=1}^n h^{-1}K(\frac{X_i-x}{h})}\\ =& \frac{\sum_{i=1}^n K(\frac{X_i-x}{h})Y_i}{\sum_{i=1}^n K(\frac{X_i-x}{h})}\end{aligned}$$

which is the same as the kernel regression estimator we defined in class (with window size $h=1$).

$$\hat{m}_{\text{ker}}(x) = \frac{1}{n}\sum_{i=1}^n W(x)Y_i = \frac{\sum_{i=1}^n K(X_i-x)Y_i}{\sum_{i=1}^n K(X_i-x)}$$

Exercise 2 Implicit Bias of Overparameterized Linear Regression

2.(a)

By taking derivative of the empirical risk $R(w)$ with respect to w, we have:

$$\begin{aligned}0 = \frac{\partial}{\partial \mathbf{w}}R(w) =& \frac{\partial}{\partial \mathbf{w}}\frac{1}{2n}\left\|\mathbf{X}\mathbf{w}-\mathbf{t}\right\|_2^2\\ =& \frac{1}{n}\left[\mathbf{X}'(\mathbf{X}\mathbf{w}-\mathbf{t})\right]\\ \Rightarrow \mathbf{X}'\mathbf{X}\mathbf{w} =& \mathbf{X}'\mathbf{t}\end{aligned}$$

which is the expression that the solution $\hat{w}$ satisfies. In underdetermined case, $\mathbf{X}'\mathbf{X}$ is not invertible, and the solution is not unique. We may, however, use e.g. the Moore-Penrose pseudoinverse to find a solution such as

$$\hat{\mathbf{w}}_{\text{Moore}-\text{Penrose}} = (\mathbf{X}'\mathbf{X})^{\dagger}\mathbf{X}'\mathbf{t}$$

2.(b)

With the gradient flow starting from $\mathbf{w}(0)=0$, we have the following ODE:

$$\mathbf{w}(t) = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t}\big(1-\exp(-t/n)\big)$$

[1]TuoruiPeng2028@u.northwestern.edu

- As $t \to \infty$, we have solution

$$\mathbf{w}(t) \to \mathbf{X}'(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t} := \mathbf{w}^*$$

- The minimum norm interpolation problem:

$$\arg\min_{\mathbf{w}} \frac{1}{2}\|\mathbf{w}\|_2^2 \quad \text{subject to} \quad \mathbf{X}\mathbf{w} = \mathbf{t}$$

its Lagrangian is

$$\mathcal{L} = \frac{1}{2}\|\mathbf{w}\|_2^2 + \lambda'(\mathbf{X}\mathbf{w} - \mathbf{t})$$

with minimizer at

$$\begin{aligned}
&\frac{\partial}{\partial \mathbf{w}}\mathcal{L} = \mathbf{w} + \mathbf{X}'\lambda = 0 \\
\Rightarrow &\mathbf{X}\mathbf{w} = \mathbf{t} = -\mathbf{X}\mathbf{X}'\lambda \\
\Rightarrow &\lambda = -(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t} \\
\Rightarrow &\mathbf{w} = \mathbf{X}'(\mathbf{X}\mathbf{X}')^{\dagger}\mathbf{t}
\end{aligned}$$

which is the same as the solution $\mathbf{w}^*$.

- Here I feel that the relation between the traditional predictor and minimum-norm predictor is similar to the relation between maximum-margin predictor and minimum-norm predictor in SVM. In this relation, the original optimize problem has a tough optimization goal (margin M in SVM, $\|Xw - t\|$ in regression), by transforming to the minimum-norm predictor, we can throw the difficult part (compute margin for SVM, determine degree of freedom for regression) into the regularization term, and solve the problem with a simpler form.

Exercise 3 Benefit of Overparameterization

The result I got is as follows. A similar result as example:

- At low degree of freedom (<7): works well, because the regression task is not complex, a relative small degree of freedom is enough to fit the data;
- At medium degree of freedom (7-70): significant overfitting and we see some terrible deviation from the true function;
- At high degree of freedom (>70): due to the overparameterization, whenever there are some data points at some region, the fitted model would be close to the data points, and thus could be close to the ture model.

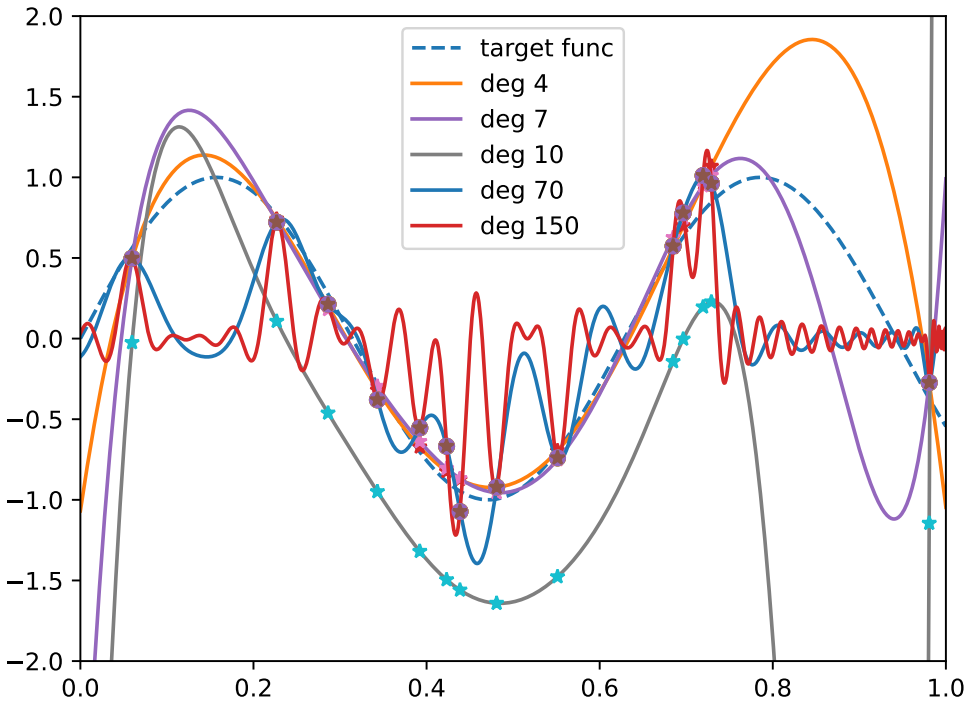

图 1: (Chebyshev) Polynomial Regression: fit v.s. degree of freedom

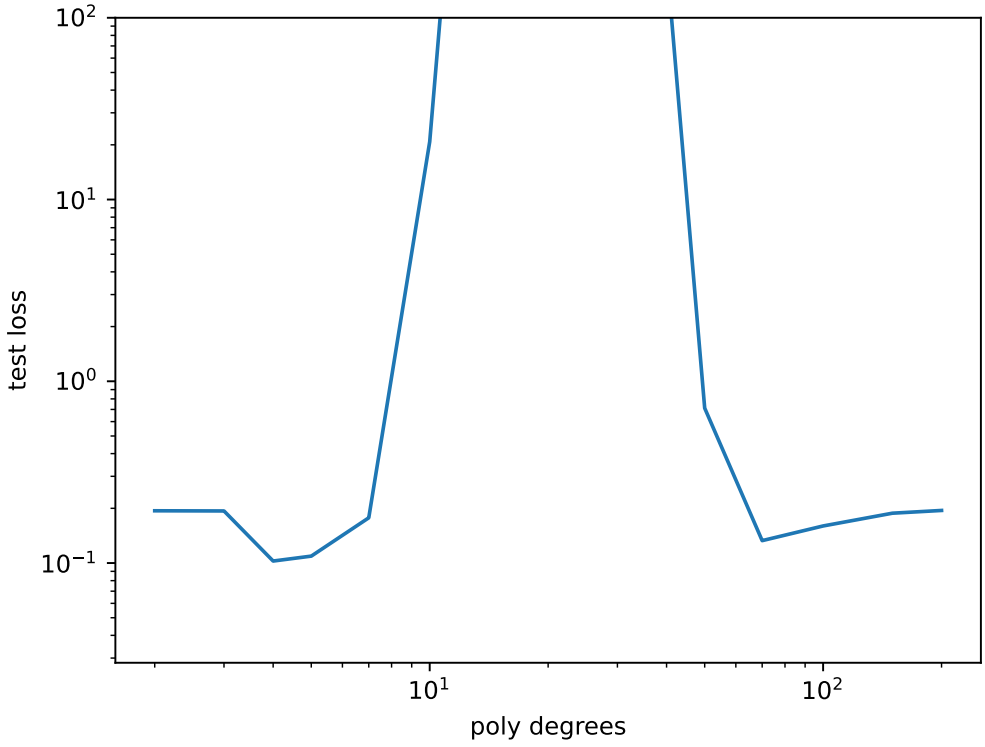

图 2: (Chebyshev) Polynomial Regression: loss v.s. degree of freedom

IEMS 402 Statistical Learning - 2025 Winter
HW3

Tuorui Peng[1]

Exercise 1 Ensemble and Bias-Variance Trade-off

1.(a) Weight Average or Prediction Average?

Note that weight average & prediction average are both linear functions of the predictions. Thus, the two are equivalent if we are using linear models (do not expending the model space), while for non-linear models like nn, the two are not equivalent (new model space).

1.(b) Bagging - Uncorrelated Models

1.(b).(i) Bias with bagging

By total expectation law, we have

$$\mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathcal{D}_i}\left[h(x;\mathcal{D}_i)|x\right]\right] = \mathbb{E}_{\mathcal{D}}\left[h(x;\mathcal{D}_i)|x\right]$$

Thus we have

$$\begin{aligned}
\text{bias} =& \mathbb{E}\left[\left(\mathbb{E}\left[\bar{h}(x;\mathcal{D})|x\right] - y_*(x)\right)^2\right] \\
=& \mathbb{E}\left[\left(\mathbb{E}\left[k^{-1}\sum_{i=1}^{k} h(x;\mathcal{D}_i)|x\right] - y_*(x)\right)^2\right] \\
=& \mathbb{E}\left[\left(\mathbb{E}\left[h(x;\mathcal{D})|x\right] - y_*(x)\right)^2\right]
\end{aligned}$$

1.(b).(ii) Variance with bagging

With uncorrelated assumption, we have

$$\begin{aligned}
\text{variance} =& \mathbb{E}\left[\left(\bar{h}(x;\mathcal{D}) - \mathbb{E}\left[\bar{h}(x;\mathcal{D})|x\right]\right)^2\right] \\
=& \mathbb{E}\left[\left(k^{-1}\sum_{i=1}^{k} h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)^2\right] \\
=& k^{-2}\sum_{i=1}^{k}\mathbb{E}\left[\left(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)^2\right] \\
&+ k^{-2}\sum_{i\neq j}\underbrace{\mathbb{E}\left[\left(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)\left(h(x;\mathcal{D}_j) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\right)\right]}_{0} \\
=& k^{-1}\sigma^2
\end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

1.(c) Bagging- General Case

1.(c).(i) Bias under Correlation

Note that in the derivation of bias, we only utilized the total expectation law, thus the bias term remains the same.

1.(c).(ii) Variance under Correlation

Turn to correlated case with $corr\big(h(x;\mathcal{D}_i), h(x;\mathcal{D}_j)\big) = \delta_{ij}(1-\rho)+\rho$, we have

$$\begin{aligned}
\text{variance} =& \mathbb{E}\left[\big(\bar{h}(x;\mathcal{D}) - \mathbb{E}\left[\bar{h}(x;\mathcal{D})|x\right]\big)^2\right] \\
=& \mathbb{E}\left[\big(k^{-1}\sum_{i=1}^{k} h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)^2\right] \\
=& k^{-2}\sum_{i=1}^{k}\mathbb{E}\left[\big(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)^2\right] \\
&+ k^{-2}\sum_{i\neq j}\mathbb{E}\left[\big(h(x;\mathcal{D}_i) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)\big(h(x;\mathcal{D}_j) - \mathbb{E}\left[h(x;\mathcal{D})|x\right]\big)\right] \\
=& k^{-1}\sigma^2 + \frac{k-1}{k}\sigma^2\rho \\
=& \sigma^2\big(\rho + \frac{1-\rho}{k}\big)
\end{aligned}$$

1.(c).(iii) Intuitions on bagging

- Generall we should have a positive ρ due to the similarity between boostrapped models, thus the variance term decreases towards $\rho\sigma^2$ as k increases. i.e. more ensembles could reduce the variance, but we would still have variance induced by the intrinsic correlation.
- $\rho = 0$ case: variance term is σ^2/k, which could reduce to 0 as k increases.
- $\rho = 1$ case: variance term is σ^2, this case bagging does not help to reduce the variance.

Exercise 2 Central Limit Theorem for Kernel Density Estimator

2.(a)

Kernel density estimator is defined as

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})$$

1. For any $q > 1$, we have

$$\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h}) - p_h(x)\right|^q\right] \begin{cases} \geq & \frac{1}{2^q}\left[\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})\right|^q\right] - \mathbb{E}\left[|p_h(x)|^q\right]\right] \\ \leq & \frac{1}{2^q}\left[\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})\right|^q\right] + \mathbb{E}\left[|p_h(x)|^q\right]\right] \end{cases}$$

in which

$$\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h})\right|^q\right] = \frac{1}{h^{q-1}}\int |K(\|v\|)|^q\, p(x+hv)\,\mathrm{d}v = \Theta(h^{-q})$$
$$\mathbb{E}\left[|p_h(x)|^q\right] = |\mathbb{E}\left[\hat{p}_h(x)\right]|^q = \Theta(n^{-q}h^{-q})$$

as a result, taking $q = 2+\delta$ we have

$$\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h}) - p_h(x)\right|^{2+\delta}\right] = \Theta(nh^{-1-\delta} + n^{-1-\delta}h^{-1-\delta}) = \Theta(nh^{-1-\delta})$$

2. Taking $q = 2$ we have

$$\big(\sum_{i=1}^{n}\sigma_i^2(x)\big)^{2+\delta} = \Theta((nh^{-1}))^{(2+\delta)/2}$$

3. Now we verify the condition for Lyaupnov CLT, we have

$$\begin{aligned}\lim_{n\to\infty}\frac{\sum_{i=1}^{n}\mathbb{E}\left[\left|\frac{1}{h}K_h(\frac{\|x-X_i\|}{h}) - p_h(x)\right|^q\right]}{(\sum_{i=1}^{n}\sigma_i^2(x))^{(2+\delta/2)}} &= \lim_{n\to\infty}\frac{\Theta(nh^{-1-\delta})}{\Theta((nh^{-1}))^{(2+\delta)/2}}\\ &= \lim_{n\to\infty}\Theta(n^{-\delta/2}h^{-\delta/2}) = 0\end{aligned}$$

thus we have the Lyaupnov CLT for KDE as

$$\frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} \xrightarrow{\mathrm{d}} N(0,1)$$

in which $s_n^2(x) = var(\hat{p}_h(x))$.

2.(b)

- Variance term:

$$s_n^2(x) = \Theta(n^{-1}h_n^{-1})$$

- Bias term:

$$\begin{aligned}|p_{h_n}(x) - p(x)| &= \left|\int K(\|v\|)\big(p(x+h_nv) - p_{x,\beta}(x+h_nv)\big)\,\mathrm{d}v\right|\\ &\sim Lh_n^{\beta}\int K(s)\,\|s\|^{\beta}\,\mathrm{d}s := ch_n^{\beta} \quad \text{generally speaking.}\end{aligned}$$

- Optimal h_n: choosing ℓ_2 risk as the criterion, we have

$$\text{error} \asymp \text{bias}^2 + \text{variance} = \Theta(h_n^{2\beta}) + \Theta(n^{-1}h_n^{-1})$$

which gives the optimal h_n as

$$h_n^* = \Theta(n^{-1/(2\beta+1)})$$

As a result, at $\beta = 2$ we have $h_n^* = \Theta(n^{-1/5})$ and

$$\frac{|p_{h_n}(x) - p(x)|}{s_n(x)} = \Theta(n^{1/2} h_n^{5/2}) = \Theta(1)$$

say the $\Theta(1)$ limit is $b(x)$, which gives

$$\frac{\hat{p}_h(x) - p(x)}{s_n(x)} = \frac{\hat{p}_h(x) - p_h(x)}{s_n(x)} + \frac{p_h(x) - p(x)}{s_n(x)} \xrightarrow{\mathrm{d}} N(0,1) + b(x) = N(b(x), 1)$$

Exercise 3 Estimating the Sobolev Ellipsoid via Spectral Methods

3.(a)

By the basis expansion, we have

$$\begin{aligned}
\text{risk} =& \mathbb{E}\left[\int_0^1 (\hat{p}(x) - p(x))^2 \,\mathrm{d}x\right] \\
=& \mathbb{E}\left[\int_0^1 \Big(\sum_{j=1}^{k} \left(\hat{\beta}_j - \beta_j\right) \phi_j(x) - \sum_{j=k+1}^{\infty} \beta_j(x)\Big)^2 \,\mathrm{d}x\right] \\
=& \sum_{j=1}^{k} \mathbb{E}\left[\left(\hat{\beta}_j - \beta_j\right)^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
=& \sum_{j=1}^{k} \mathbb{E}\left[\left(n^{-1} \sum_{i=1}^{n} \phi_j(X_i) - \beta_j\right)^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2
\end{aligned}$$

Now note that we have

$$\mathbb{E}\left[\phi_j(X_i)\right] = \int_0^1 \phi_j(x) p(x) \,\mathrm{d}x = \beta_j$$

thus we have

$$\begin{aligned}
\text{risk} =& \sum_{j=1}^{k} \mathbb{E}\left[\left(n^{-1} \sum_{i=1}^{n} \phi_j(X_i) - \beta_j\right)^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
=& \sum_{j=1}^{k} \left[\frac{1}{n^2} \sum_{i=1}^{n} \mathbb{E}\left[\phi_j(X_i)^2\right] - 2\beta_j \frac{1}{n} \sum_{i=1}^{n} \beta_j + \beta_j^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
\leq& \sum_{j=1}^{k} \left[\frac{C^2}{n} - \beta_j^2\right] + \sum_{j=k+1}^{\infty} \beta_j^2 \\
\leq& \frac{ck}{n} + \sum_{j=k+1}^{\infty} \beta_j^2.
\end{aligned}$$

3.(b)

We have for Sobolev ellipsoid $E(m, L)$ that

$$k^{2m} \sum_{j=k+1}^{\infty} \beta_j^2 \leq \sum_{j=1}^{k} \beta_j^2 j^{2m} + \sum_{j=k+1}^{\infty} \beta_j^2 j^{2m} < L^2 \Rightarrow \sum_{j=k+1}^{\infty} \beta_j^2 < \frac{L^2}{k^{2m}}$$

as a result we have

$$\text{risk} \leq \frac{ck}{n} + \frac{L^2}{k^{2m}} \lesssim \frac{k}{n} + \left(\frac{1}{k}\right)^{2m}$$

The optimal k is obtained at $k \asymp n^{1/(2m+1)}$, which gives the optimal risk $\lesssim n^{-2m/(2m+1)}$.

IEMS 402 Statistical Learning - 2025 Winter

HW4

Tuorui Peng[1]

Exercise 1 Estimating the Derivatives via Kernel Smoothing

- Bias term:

$$\begin{aligned}
\left|d_n(x) - p'(x)\right| =& \left|\int_{-1}^{1} \frac{1}{h^2} K\left(\frac{X_i - x}{h}\right) p(x)\,\mathrm{d}x - p'(x)\right| \\
=& \frac{1}{h}\left|\int K(\|v\|)\big(p(x+hv) - vp'(x)\big)\,\mathrm{d}v\right| \\
=& \frac{1}{h}\left|\int K(\|v\|)\big(p(x+hv) - vp'_{x,\beta}(x+hv) + vp'_{x,\beta}(x+hv) - vp'(x)\big)\,\mathrm{d}v\right| \\
\leq& \frac{1}{h}\int K(\|v\|)\left|p(x+hv) - vp'_{x,\beta}(x+hv)\right|\,\mathrm{d}v + \frac{1}{h}\int K(\|v\|)\left|vp'_{x,\beta}(x+hv) - vp'(x)\right|\,\mathrm{d}v \\
\lesssim& Lh^{\beta-1}\int K(\|v\|)\,|v|\,\mathrm{d}v
\end{aligned}$$

  in which the last inequality is because $p(\,\cdot\,) - p_{x,\beta}(\,\cdot\,)$ is still a β-Hölder function.

$$\begin{aligned}
var(d_n(x)) \leq& \frac{1}{nh^4}\int K^2(\frac{X_i - x}{h})p(x)\,\mathrm{d}x \\
=& \frac{1}{nh^3}\int K^2(v)p(x+hv)\,\mathrm{d}v \\
\lesssim& \frac{1}{nh^3}\sup_{x\in[-1,1]} p(x)\int K^2(v)\,\mathrm{d}v
\end{aligned}$$

  where the last inequality is because only the first order term of $p(\,\cdot\,)$ gives non-zero kernel integration.

Put together, we have

$$\mathrm{MSE} \lesssim h^{2(\beta-1)} + \frac{1}{nh^3}$$

Optimal h_n is at $h_n \asymp n^{-\frac{1}{2\beta+1}}$, and the optimal MSE is $n^{-\frac{2(\beta-1)}{2\beta+1}}$.

Exercise 2 An average treatment effect estimator

2.(a)

In completely randomized experiment, we can see that

$$\begin{aligned}
\mathbb{E}\left[Y_i(a)1\{A_i = a\}\right] =& \frac{1}{2}\mathbb{E}\left[Y_i(a)1\{A_i = a\}|A_i = 0\right] + \frac{1}{2}\mathbb{E}\left[Y_i(a)1\{A_i = a\}|A_i = 1\right] \\
=& \begin{cases} \frac{1}{2}\mathbb{E}\left[Y_i(0)|A_i = 0\right], & \text{if } a = 0 \\ \frac{1}{2}\mathbb{E}\left[Y_i(1)|A_i = 1\right], & \text{if } a = 1 \end{cases} \\
=& \frac{1}{2}\mathbb{E}\left[Y_i(a)\right] = \frac{1}{2}\mathbb{E}\left[Y(a)\right]
\end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

and we have ATE being

$$\tau = \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right] = 2\mathbb{E}\left[Y(1)1\{A=1\}\right] - 2\mathbb{E}\left[Y(0)1\{A=0\}\right].$$

2.(b)

Note that we can write $\hat{\tau}_n$ as

$$\begin{aligned}
\hat{\tau}_n =& \frac{1}{n}\big[2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}\big] \\
\mathbb{E}\left[2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}\right] =& \tau \\
var(2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}) =& var(\mathbb{E}\left[2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}|A\right]) \\
&+ \mathbb{E}\left[var(2Y_i(1)1\{A_i=1\} - 2Y_i(0)1\{A_i=0\}|A)\right] \\
=& (\mathbb{E}\left[Y(1)\right] + \mathbb{E}\left[Y(0)\right])^2 + 2\big(var(Y(1)) + var(Y(0))\big)
\end{aligned}$$

By CLT

$$\sqrt{n}(\hat{\tau}_n - \tau) \xrightarrow{\text{d}} N(0, (\mu_1+\mu_0)^2 + 2(\sigma_1^2+\sigma_0^2)), \quad \mu_a = \mathbb{E}\left[Y(a)\right], \sigma_a^2 = var(Y(a))$$

2.(c)

Note that for completely randomized experiment, we have $|S_1| = n - |S_0| \sim \text{Binomial}(n, \frac{1}{2})$. Thus we get (take $a=1$ example)

$$\begin{cases} \sqrt{n}(1 - 2\left|S_1\right|/n) \xrightarrow{\text{d}} N(0,1) \\ 2\left|S_1\right|/n \xrightarrow{\text{p}} 1 \end{cases} \overset{\text{Slutsky}}{\Rightarrow} \sqrt{n}(\frac{n}{2\left|S_1\right|} - 1) \xrightarrow{\text{d}} N(0,1)$$

and similarly for $a=0$ we have $\sqrt{n}(\frac{n}{2\left|S_0\right|} - 1) \xrightarrow{\text{d}} N(0,1)$.

2.(d)

We have

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau) =& \frac{\sqrt{n}}{\sqrt{|S_1|}}\sqrt{|S_1|}\big(\frac{1}{|S_1|}\sum_i (Y_i(1) - \mathbb{E}\left[Y(1)\right])1\{A_i=1\}\big) \\
&+ \frac{\sqrt{n}}{\sqrt{|S_0|}}\sqrt{|S_0|}\big(\frac{1}{|S_0|}\sum_i (Y_i(0) - \mathbb{E}\left[Y(0)\right])1\{A_i=0\}\big)
\end{aligned}$$

For now we treat A as given, and we have

$$\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau)\big|A \xrightarrow{\text{d}} \frac{\sqrt{n}}{\sqrt{|S_1|}}N(0,\sigma_1^2) + \frac{\sqrt{n}}{\sqrt{|S_0|}}N(0,\sigma_0^2)$$

On the other hand we notice that we already have

$$\frac{\sqrt{n}}{\sqrt{|S_1|}} \xrightarrow{\text{p}} \sqrt{2}, \quad \frac{\sqrt{n}}{\sqrt{|S_0|}} \xrightarrow{\text{p}} \sqrt{2}, \quad cov\left(\frac{n}{|S_1|}, \frac{n}{|S_0|}\right) = -1$$

Thus by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau) \xrightarrow{\text{d}} N(0, \sigma^2_{\text{norm}})$$

where

$$\sigma^2_{\text{norm}} = 2\sigma_1^2 + 2\sigma_0^2$$

2.(e)

It seems that based on our results up to now, the conlusion would be that:

$$\begin{aligned}\sqrt{n}(\hat{\tau}_n - \tau) &\xrightarrow{\text{d}} N(0, \sigma^2), \quad \sigma^2 = (\tau_1 + \tau_0)^2 + 2(\sigma_1^2 + \sigma_0^2)\\ \sqrt{n}(\hat{\tau}_n^{\text{norm}} - \tau) &\xrightarrow{\text{d}} N(0, \sigma^2_{\text{norm}}), \quad \sigma^2_{\text{norm}} = 2(\sigma_1^2 + \sigma_0^2)\end{aligned}$$

so we have $\sigma^2 > \sigma^2_{\text{norm}}$ as long as $\tau_1 + \tau_0 > 0$.

Exercise 3 A weighted average treatment effect estimator

3.(a)

With the covariate X involved, we have

$$\begin{aligned}\tau =& \mathbb{E}\left[Y(1)\right] - \mathbb{E}\left[Y(0)\right]\\ =& \mathbb{E}\left[\mathbb{E}\left[Y(1)\big(1\{A=1\} + 1\{A=0\}\big)\big|X=x\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y(0)\big(1\{A=1\} + 1\{A=0\}\big)\big|X=x\right]\right]\end{aligned}$$

now note that since $(Y(1), Y(0)) \perp A|X$, we have

$$\begin{aligned}\mathbb{E}\left[Y(A)1\{A=1\}|X=x\right] =& Y(1)e(x)\\ \mathbb{E}\left[Y(A)1\{A=0\}|X=x\right] =& Y(0)(1-e(x))\end{aligned}$$

substitute this back to the above equation, we have

$$\begin{aligned}\tau =& \mathbb{E}\left[\frac{Y(A)1\{A=1\}}{e(x)}\big(1\{A=1\} + 1\{A=0\}\big) - \frac{Y(A)1\{A=0\}}{1-e(x)}\big(1\{A=1\} + 1\{A=0\}\big)\Big|X=x\right]\\ =& \mathbb{E}\left[\frac{Y(A)1\{A=1\}}{e(x)} - \frac{Y(A)1\{A=0\}}{1-e(x)}\Big|X=x\right]\\ =& \mathbb{E}\left[\frac{Y(A)1\{A=1\}}{e(X)}\right] - \mathbb{E}\left[\frac{Y(A)1\{A=0\}}{1-e(X)}\right]\end{aligned}$$

3.(b)

By CLT, the propensity weighted estimator

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i 1\{A_i=1\}}{e(X_i)} - \frac{Y_i 1\{A_i=0\}}{1-e(X_i)}\right] \xrightarrow{\text{d}} N(\tau, n\sigma^2_{\text{ps}})$$

where the variance is computed as follows:

- Prepare for the calculation:

$$\begin{aligned}
\mathbb{E}\left[(1\{A_i=1\}-e(X_i))Y_i|X_i=x\right] =&0\\
var((1\{A_i=1\}-e(X_i))Y_i|X_i=x) =&e(x)(1-e(x))\mathbb{E}\left[Y_i^2|X_i=x\right]\\
=&e(x)(1-e(x))\mathbb{E}\left[(Y(1)1\{A=1\}+Y(0)1\{A=0\})^2|X=x\right]\\
=&e(x)(1-e(x))\big(e(x)v_2(x,1)^2+(1-e(x))v_2(x,0)^2\big)
\end{aligned}$$

$$\begin{aligned}
\sigma^2_{\text{ps}} =&var(\frac{Y_i1\{A_i=1\}}{e(X_i)}-\frac{Y_i1\{A_i=0\}}{1-e(X_i)})\\
=&var(\frac{(1\{A_i=1\}-e(X_i))Y_i}{e(X_i)(1-e(X_i))})\\
=&\mathbb{E}\left[var(\frac{(1\{A_i=1\}-e(X_i))Y_i}{e(X_i)(1-e(X_i))}\big|X_i=x)\right]+var\left(\mathbb{E}\left[\frac{(1\{A_i=1\}-e(X_i))Y_i}{e(X_i)(1-e(X_i))}\big|X_i=x\right]\right)\\
=&\mathbb{E}\left[\frac{\big(e(X)v_2(X,1)^2+(1-e(X))v_2(X,0)^2\big)}{e(X)(1-e(X))}\right]\\
=&\mathbb{E}\left[\frac{v_2(X,1)^2}{1-e(X)}\right]+\mathbb{E}\left[\frac{v_2(X,0)^2}{e(X)}\right]
\end{aligned}$$

3.(c)

Note that, given $X=x$, we have

$$\sigma^2_{\text{ps},x}=\left(\frac{v_2(x,1)^2}{1-e(x)}+\frac{v_2(x,0)^2}{e(x)}\right)\cdot(1-e(x)+e(x))\overset{\text{Cauchy-Schwarz}}{\geq}(v_2(x,1)+v_2(x,0))^2$$

which takes the minimum when

$$\frac{v_2(x,1)}{1-e(x)}=\frac{v_2(x,0)}{e(x)}\Rightarrow e(x)=\frac{v_2(x,0)}{v_2(x,0)+v_2(x,1)}$$

which is the optimal propensity score chosen to minimize the variance of the propensity score weighted estimator.

One sentence intuition: the choice of propensity score should balance variance within two groups w.r.t. the covariate X, i.e. similar to the idea to avoiding simpson's paradox. Thus this method can lead to significant improvement over the naive randomized experiment estimator when we do have a good covariate X to work with.

Exercise 4 Logistic regression

For logistic regression, we have

$$\begin{aligned}
\ell(\theta):=\text{ log-likelihood }(\theta) =&\frac{1}{n}\sum_{i=1}^n Y_i\log\pi_\theta(X_i)+(1-Y_i)\log(1-\pi_\theta(X_i)),\quad \pi_\theta(X)=\frac{1}{1+\exp(-\theta'X)}\\
\frac{\partial^2\ell}{\partial\theta\partial\theta'} =&\frac{1}{n}\sum_{i=1}^n\frac{\exp(-\theta'X_i)}{(1+\exp(-\theta'X_i))^2}X_iX_i'\\
I(\theta):=&\mathbb{E}\left[\frac{\exp(-\theta'X_i)}{(1+\exp(-\theta'X_i))^2}X_iX_i'\right]
\end{aligned}$$

We have convergence of $\hat{\theta}_n = \arg\max_{\theta} \ell(\theta)$ that:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathrm{d}} N(0, I(\theta)^{-1})$$

IEMS 402 Statistical Learning - 2025 Winter

HW4

Tuorui Peng[1]

Exercise 1 Le Cam One-step estimators

For convenience, we label $l_n(\theta) = \frac{1}{n} L_n(\theta)$.

Solving the first order equation is equivalent to solving the following equation: (I guess there is a typo in the question setting, should be $\nabla l_n(\hat{\theta}_n) + \nabla^2 l_n(\hat{\theta}_n)\delta_n = 0$)

$$\delta_n = -\left(\nabla^2 l_n(\hat{\theta}_n)\right)^{-1} \nabla l_n(\hat{\theta}_n)$$

in which note that $\hat{\theta}_n - \theta_0 = O(n^{-1/2})$.

Then we have

- $$\nabla l_n(\theta_0) \xrightarrow{\mathrm{d}} \mathcal{N}(0, var(\nabla \ell(\theta_0, X))/n) = \mathcal{N}(0, I(\theta_0)/n)$$

- $$\nabla^2 l_n(\theta_0) = \nabla^2 \mathbb{E}\left[\ell(\theta_0, X)\right] + O(n^{-1/2})$$

- $$\begin{aligned} \nabla^2 l_n(\hat{\theta}_n) - \nabla^2 l_n(\theta_0) =& \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(\hat{\theta}_n, X_i) - \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(\theta_0, X_i) \\ \leq& \frac{1}{n}\sum_{i=1}^n M(X_i) \left|\hat{\theta}_n - \theta_0\right| \\ =& \left|\hat{\theta}_n - \theta_0\right| \left(\mathbb{E}\left[M(X)\right] + O(n^{-1/2})\right) \\ =& O(n^{-1/2}) \end{aligned}$$

- $$\begin{aligned} \nabla l_n(\hat{\theta}_n) - \nabla l_n(\theta_0) =& \frac{1}{n}\sum_{i=1}^n \nabla \ell(\hat{\theta}_n, X_i) - \frac{1}{n}\sum_{i=1}^n \nabla \ell(\theta_0, X_i) \\ =& \frac{1}{n}\sum_{i=1}^n \nabla^2 \ell(\theta_0, X_i)(\hat{\theta}_n - \theta_0) + o(\hat{\theta}_n - \theta_0) \\ =& (\nabla^2 l_n(\theta_0) + O(n^{-1/2}))(\hat{\theta}_n - \theta_0) + o(n^{-1/2}) \\ =& \nabla^2 l_n(\theta_0)(\hat{\theta}_n - \theta_0) + o(n^{-1/2}) \\ =& \nabla^2 \mathbb{E}\left[\ell(\theta_0, X)\right](\hat{\theta}_n - \theta_0) + o(n^{-1/2}) \end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

Now we have

$$\begin{aligned}\delta_n =& -(\nabla^2 l_n(\hat{\theta}_n))^{-1}\nabla l_n(\hat{\theta}_n)\\ =& -(\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right]+O(n^{-1/2}))^{-1}(\mathcal{N}(0,I(\theta_0)/n)+\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right](\hat{\theta}_n-\theta_0)+o(n^{-1/2}))\\ =& -\left(\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right]^{-1}+O(n^{-1/2})\right)(\mathcal{N}(0,I(\theta_0)/n)+\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right](\hat{\theta}_n-\theta_0)+o(n^{-1/2}))\\ \xrightarrow{\mathrm{d}}&\theta_0-\hat{\theta}_n+\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right]\mathcal{N}(0,I(\theta_0)/n)+o(n^{-1/2})\end{aligned}$$

then

$$\sqrt{n}(\bar{\theta}_n-\theta_0)=\sqrt{n}(\hat{\theta}_n+\delta_n)\xrightarrow{\mathrm{d}}\mathcal{N}(\theta_0,(\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right])^{-1}I(\theta_0)(\nabla^2\mathbb{E}\left[\ell(\theta_0,X)\right])^{-1})$$

Exercise 2 Sub-Gaussianity of bounded R.V.s

2.(a)

Consider $W:=Y-(a+b)/2$, for which we notice that $|W|\leq(b-a)/2$ and $var(W)=var(Y)$. Now we have

$$\begin{aligned}var(Y)=&var(W)\\ =&\mathbb{E}\left[W^2\right]-\mathbb{E}\left[W\right]^2\\ \leq&\mathbb{E}\left[W^2\right]\\ \leq&\frac{(b-a)^2}{4}\end{aligned}$$

2.(b)

We have

$$\phi'(\lambda)=\frac{\mathbb{E}_P\left[Xe^{\lambda X}\right]}{\mathbb{E}_P\left[e^{\lambda X}\right]}=\mathbb{E}_P\left[X\frac{e^{\lambda X}}{\mathbb{E}_P\left[e^{\lambda X}\right]}\right]=\mathbb{E}_{Q_\lambda}\left[X\right]$$

and

$$\begin{aligned}\phi''(\lambda)=&\frac{\mathbb{E}_P\left[X^2e^{\lambda X}\right]\mathbb{E}_P\left[e^{\lambda X}\right]-\mathbb{E}_P\left[Xe^{\lambda X}\right]^2}{\mathbb{E}_P\left[e^{\lambda X}\right]^2}\\ =&\mathbb{E}_P\left[X^2\frac{e^{\lambda X}}{\mathbb{E}_P\left[e^{\lambda X}\right]}\right]-\left(\mathbb{E}_P\left[X\frac{e^{\lambda X}}{\mathbb{E}_P\left[e^{\lambda X}\right]}\right]^2\right)\\ =&\mathbb{E}_{Q_\lambda}\left[X^2\right]-\left(\mathbb{E}_{Q_\lambda}\left[X\right]\right)^2\\ =&var_{Q_\lambda}(X)\end{aligned}$$

2.(c)

In this part we consider relabel $\phi(\lambda)=\log\mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right]$.

Note that if $X\in[a,b]$, then we have

$$\phi''(\lambda)=var_{Q_\lambda}(X)\leq\frac{(b-a)^2}{4}$$

on the other hand we notice that $\phi(0) = 0$ and $\phi'(0) = 0$

$$\begin{aligned}\phi'(\lambda) =& \phi'(0) + \int_0^\lambda \phi''(t)dt \leq \frac{(b-a)^2}{4}\lambda \\ \phi(\lambda) =& \phi(0) + \int_0^\lambda \phi'(t)dt \leq \frac{(b-a)^2}{8}\lambda^2\end{aligned}$$

Together we have

$$\phi(\lambda) = \log \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq \frac{(b-a)^2}{8}\lambda^2 \Rightarrow \mathbb{E}\left[e^{\lambda(X-\mathbb{E}[X])}\right] \leq e^{(b-a)^2\lambda^2/8}$$

i.e. X is sub-Gaussian with parameter $(b-a)^2/4$.

Exercise 3 Concentration inequalities

Note that by

3.(a)

By taylor expansion we have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda X}\right] =& \mathbb{E}\left[1 + \lambda X + \frac{\lambda^2 X^2}{2} + \sum_{k=3}^{\infty}\frac{\lambda^k X^k}{k!}\right] \\ \leq& 1 + 0 + \mathbb{E}\left[\frac{X^2}{c^2}\sum_{k=2}^{\infty}\frac{\lambda^k c^k}{k!}\right] \\ =& 1 + \frac{\sigma^2}{c^2}(e^{\lambda c} - 1 - \lambda c) \\ \leq& \exp\left[\frac{\sigma^2}{c^2}(e^{\lambda c} - 1 - \lambda c)\right]\end{aligned}$$

3.(b) Bennett's inequality

Using the result from the previous part, we have

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) =& \mathbb{P}\left(e^{\lambda\sum_{i=1}^n X_i} \geq e^{\lambda t}\right) \\ \leq& \frac{\mathbb{E}\left[e^{\lambda\sum_{i=1}^n X_i}\right]}{e^{\lambda t}} \\ \leq& \exp\left[\frac{\sum_{i=1}^n \sigma_i^2}{c^2}(e^{\lambda c} - 1 - \lambda c) - \lambda t\right]\end{aligned}$$

Optimizing the right hand side with respect to λ, we have optimal $\lambda_* = \frac{1}{c}\log\frac{ct}{n\sigma^2}$. Substituting this back we have

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \exp\left[\frac{n\sigma^2}{c^2}(\frac{ct}{n\sigma^2} - 1 - \log\frac{ct}{n\sigma^2}) - \frac{t}{c}\log\frac{ct}{n\sigma^2}\right] = \exp\left[-\frac{n\sigma^2}{c^2}h(\frac{ct}{n\sigma^2})\right]$$

3.(c) Bernstein's inequality

It suffices to prove for one sided part (cuz we can apply the whole part to $-X_i$).

$$h(u) = (1+u)\log(1+u) - u$$

We have

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq nt\right) \leq \exp\left[-\frac{n\sigma^2}{c^2} h(\frac{ct}{\sigma^2})\right]$$

so it suffice to prove the following inequality:

$$h(u) \geq \frac{u^2}{2+2u/3}, \quad u \geq 0$$

which is trivial by taking derivative.

$$\begin{aligned}
g(u) :=& (2+2u/3)h(u) - u^2, \quad g(0) = 0 \\
g'(u) =& \frac{4}{3}\log(u+1) + \frac{2}{3}(2(u+1)\log(u+1) - u) - 2u, \quad g'(0) = 0 \\
g''(u) =& \frac{4}{3(u+1)} + \frac{4}{3}\log(u+1) - \frac{4}{3} \geq 0
\end{aligned}$$

With the above, substitute $u = \dfrac{ct}{\sigma^2}$ and we can obtain the desired result.

3.(d)

Berstein' inequality is stronger than Hoeffding's inequality for small t, to be specific, when

$$ct \lesssim \sigma^2, \quad i.e. \quad t \lesssim \frac{\sigma^2}{c}$$

Exercise 4 Application of Concentration Inequalities

It suffices to lower bound the covering number of $\{0,1\}^n$.

WLOG assume $n/4$ is an integer, otherwise we apply the following argument to $\lfloor n/4 \rfloor$.

Assume we have a $n/4$ minimum covering set $\mathcal{S} = \{z_1, \ldots, z_S\}$. For each $i \in [S]$, there are $\sum_{i=0}^{n/4} \binom{n}{n/4}$ points that are in $\mathcal{B}_H(z_i, n/4)$. And this covering sould cover all 2^n points, so we have

$$S\sum_{i=0}^{n/4}\binom{n}{n/4} \geq 2^n \Rightarrow \frac{1}{S} \leq \frac{2^n}{\sum_{i=0}^{n/4}\binom{n}{n/4}} \asymp \mathbb{P}\left(\text{Binom}(n,1/2) < n/4\right) \leq \exp\left[-\frac{n}{8}\right]$$

As a result, we have

$$S \geq \exp\left[\frac{n}{8}\right]$$

Next note that we have relation between covering number and packing number, thus gives that

$$M(\{0,1\}^n, n/4) \geq N(\{0,1\}^n, n/4) = S \geq \exp\left[\frac{n}{8}\right]$$

this packing set is the set A desired that satisfies $|A| \leq e^{n/8}$, while any $x, y \in A$ satisfies $\|x-y\|_{\text{Hamming}} \geq n/4$.

IEMS 402 Statistical Learning - 2025 Winter
HW4

Tuorui Peng[1]

Exercise 1 Fast Rate Generalization Error in the Realizable Setting

For 0-1 loss, we notice that: if $L(h^*) = 0$ is reachable, then it means that $\exists h_* \in \mathcal{H}$ s.t. $y^{(i)} = h_*(x^{(i)})$ a.s. (i.e. h_* is the underlying truth, and the data generation $y|x$ has no randomness in the sense of a.s.). Then we have

$$\hat{L}(h) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((x^{(i)}, y^{(i)}), h\right) = \frac{1}{n}\sum_{i=1}^{n} \ell\left((x^{(i)}, h_*(x^{(i)}), h)\right)$$

can reach $\hat{L}(h) = 0$ in some non-empty set $H \subset \mathcal{H}$ (because we would have relation $h_* \in \{h^*\} \subseteq H \subseteq \mathcal{H}$). So the event of interest is that $\hat{h} \in H \backslash \{h^*\}$

For any $h \in \mathcal{H}$ s.t. $L(h) \geq t$, we have:

$$\begin{aligned}
\mathbb{P}\left(\hat{L}(h) = 0\right) =& \mathbb{P}\left(\forall i : \ \ell\big((x^{(i)}, y^{(i)}), h\big) = 0\right) \\
=& \mathbb{P}\left(\ell\big((x, h_*(x)), h\big) = 0\right)^n \\
=& \big(1 - \mathbb{E}\left[\ell\big((x, h_*(x)), h\big)\right]\big)^n \\
\leq& (1-t)^n \leq \exp\left(-nt\right)
\end{aligned}$$

So we have

$$\begin{aligned}
\mathbb{P}\left(L(\hat{h}) \geq t\right) =& \mathbb{P}\left(\exists h : \ L(h) \geq t, \ \hat{L}(h) = 0\right) \\
\leq& \sum_{h:\, L(h) \geq t} \mathbb{P}\left(\hat{L}(h) = 0\right) \\
\leq& |\mathcal{H}| \cdot \exp\left(-nt\right)
\end{aligned}$$

Now set $\delta = |\mathcal{H}| \cdot \exp\left(-nt\right)$ and we have

$$L(\hat{h}) \leq \frac{\log|\mathcal{H}| + \log\frac{1}{\delta}}{n}$$

Exercise 2 Generalization Error near Interpolate

2.(a)

Note that for any fixed h, we have $\mathbb{E}\left[\hat{L}_n(h)\right] = L(h) = \mathbb{E}\left[\ell(h)\right]$ and we have

$$var(\ell(h)) = \mathbb{E}\left[\ell(h)^2\right] - \mathbb{E}\left[\ell(h)\right]^2 \leq \mathbb{E}\left[\ell(h)\right] = L(h)$$

since $\ell(h) \in [0, 1]$. Now by Bernstein's inequality we proved in the previous homework, we have

$$\mathbb{P}\left(\hat{L}(h) - L(h) \geq \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(var(\ell(h)) + \varepsilon/3)}\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(L(h) + \varepsilon/3)}\right)$$

[1] TuoruiPeng2028@u.northwestern.edu

2.(b)

By the same argument we also have inequality for the other side:

$$\mathbb{P}\left(\hat{L}(h) - L(h) \leq -\varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(L(h) + \varepsilon/3)}\right)$$

Substitute $\hat{L}_n(h) \mapsto \hat{L}_n(h) - (L(h) - \varepsilon(h) - \varepsilon)$ and we have

$$\mathbb{P}\left(\hat{L}(h) \leq \varepsilon(h)\right) \leq \mathbb{P}\left(\hat{L}(h) - L(h) \leq -\varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon(h) + \varepsilon + \varepsilon/3)}\right) = \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon(h) + 4\varepsilon/3)}\right)$$

2.(c)

Note that by (b), we have for all h satisfying $L(h) \geq L(h^*) + 2\varepsilon$ that

$$\mathbb{P}\left(\hat{L}(h) \leq L(h^*) + \varepsilon\right) \leq \exp\left(-\frac{n\varepsilon^2}{2(\varepsilon + 4\varepsilon/3)}\right) = \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right)$$

So we have

$$\begin{aligned}
\mathbb{P}\left(\hat{L}_n(\hat{h}_n) - L(h^*) \geq 2\varepsilon\right) \leq & \mathbb{P}\left(\exists h:\, L(h) \geq L(h^*) + 2\varepsilon,\, \hat{L}(h) \leq L(h^*) + \varepsilon\right) \\
\leq & \sum_{h: L(h) \geq L(h^*) + 2\varepsilon} \mathbb{P}\left(\hat{L}(h) \leq L(h^*) + \varepsilon\right) \\
\leq & |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right)
\end{aligned}$$

On the other hand, by Hoeffding's inequality we have

$$\mathbb{P}\left(\left|\hat{L}_n(\hat{h}_n) - L(\hat{h}_n)\right| \geq \varepsilon\right) \leq \mathbb{P}\left(\sup_{h \in \mathcal{H}} \left|\hat{L}_n(h) - L(h)\right| \geq \varepsilon\right) \leq |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2}\right)$$

Together, we have

$$\begin{aligned}
\mathbb{P}\left(L(\hat{h}_n) - L(h^*) \leq \varepsilon\right) \geq & \mathbb{P}\left(\hat{L}_n(\hat{h}_n) - L(h^*) \geq 2\varepsilon, L(\hat{h}_n) - \hat{L}_n(\hat{h}_n) \leq \varepsilon\right) \\
\geq & 1 - |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right) - |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2}\right)
\end{aligned}$$

Note that for large ε, the above probability bound is dominated by the second term, so we consider setting $\delta = |\mathcal{H}| \exp\left(-\frac{n\varepsilon^2}{2(L(h^*) + 7\varepsilon/3)}\right)$ and we have

$$w.p. \geq 1 - \delta, \quad L(\hat{h}_n) - L(h^*) \lesssim \sqrt{\frac{L(h^*) \log(|\mathcal{H}|/\delta)}{n}} + \frac{\log(|\mathcal{H}|/\delta)}{n}.$$

2.(d)

The naive approach based on Hoeffding's inequality would give us

$$L(\hat{h}_n) - L(h^*) \lesssim \sqrt{\frac{\log|H| + \log\frac{1}{\delta}}{n}}$$

By comparing the two bounds, we notice that: when we have $L(h^*) \ll 1/\log(|\mathcal{H}|/\delta)$ the bound is dominated by $O(n^{-1})$, which is a stronger bound than the naive approach $O(n^{-1/2})$, leading to a faster rate of convergence.

Exercise 3 Random Matrix

3.(a)

Denote the SVD of A as

$$A = U\Sigma V^T$$

and thus

$$\|A\|_{\text{op}} = \sigma_1 = \max_{u\in S^{m-1}, v\in S^{n-1}} \max_{u\in S^{m-1}, v\in S^{n-1}} u'Av = \max_{u\in S^{m-1}, v\in S^{n-1}} \langle Au, v\rangle$$

by SVD.

3.(b)

(For this part I feel that the best I can do is $\sim \dfrac{1}{1-2\varepsilon}$ in stead of $\sim \dfrac{1}{(1-\varepsilon)^2}$)

With ε-net of S^{m-1} and S^{n-1} denoted $\mathcal{N}_1$ and $\mathcal{N}_2$ respectively. And denote the maximizer of $\langle Au, v'\rangle$ as (u^*, v^*), we have some $u_1 \in \mathcal{N}_1$ and $v_1 \in \mathcal{N}_2$ s.t.

$$\|u_1 - u^*\| \le \varepsilon, \quad \|v_1 - v^*\| \le \varepsilon$$

Then we have for $u^* \in S^{m-1}, v^* \in S^{n-1}$ and their corresponding u_1, v_1 we have

$$\begin{aligned}
\|A\|_{\text{op}} &= \langle Au, v\rangle \\
&= \langle A(u - u_1 + u_1), (v - v_1 + v_1)\rangle \\
&= \langle Au_1, v_1\rangle + \langle A(u - u_1), v_1\rangle + \langle Au, (v - v_1)\rangle \\
&\le \max_{u\in\mathcal{U}, v\in\mathcal{V}} \langle Au, v\rangle + \|A\|_{\text{op}}\,\varepsilon + \|A\|_{\text{op}}\,\varepsilon \\
\Rightarrow \|A\|_{\text{op}} &\le \frac{1}{1-2\varepsilon} \max_{u\in\mathcal{U}, v\in\mathcal{V}} \langle Au, v\rangle
\end{aligned}$$

3.(c)

Note that for any $u \in S^{m-1}, v \in S^{n-1}$, we have

$$\langle Au, v\rangle \sim N(0, 1)$$

And for ε covering set of S^{m-1} and S^{n-1}, we have

$$|\mathcal{U}| \le |\mathcal{N}(B_m(1)), \varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^m$$
$$|\mathcal{V}| \le |\mathcal{N}(B_n(1)), \varepsilon| \le \left(1 + \frac{2}{\varepsilon}\right)^n$$

We have

$$\begin{aligned}\mathbb{E}\left[\|A\|_{\text{op}}\right] \leq & \frac{1}{1-2\varepsilon}\mathbb{E}\left[\max_{u\in\mathcal{U},v\in\mathcal{V}}\langle Au,v\rangle\right]\\ \lesssim & \frac{1}{1-2\varepsilon}\sqrt{\log|U|+\log|V|}\\ \lesssim & \sqrt{m}+\sqrt{n}\end{aligned}$$

where the last step can be done by taking some small constant ε.

3.(d)

For S^{n-1}, we construct a $\sqrt{2}$-net $\mathcal{N}$ formed by all unit vectors. This can be easily verified to be a covering by noticing that for $\sum_{i=1}^n x_i^2 \leq 1$ with WLOG $x_1 \geq 0$ we have

$$(1-x_1^2)+\sum_{i=2}^n x_i^2 \leq 2$$

Using this method we obtain covering for S^{m-1} and S^{n-1}respectively, and we have their size bounded by $(1+2/\sqrt{2})^m$ and $(1+2/\sqrt{2})^n$ respectively.

With this construction, we have

$$\langle u_i,u_j\rangle=\delta_{ij},\quad \langle v_i,v_j\rangle=\delta_{ij},\quad u_i,u_j\in\mathcal{N}_1;\ v_i,v_j\in\mathcal{N}_2$$

and thus for any $u_1,u_2\in\mathcal{N}_1$ and $v_1,v_2\in\mathcal{N}_2$ we have

$$\begin{aligned}cov\left(\langle Au_1,v_1\rangle,\langle Au_2,v_2\rangle\right) =& cov(\sum_{i=1}^m u_{1i}v_{1j}A_{ij},\sum_{i=1}^n u_{2i}v_{2j}A_{ij})\\ =&\sum_{i=1}^m\sum_{j=1}^n u_{1i}v_{1j}u_{2i}v_{2j}\\ =&\langle u_1,u_2\rangle\langle v_1,v_2\rangle\\ =&0\end{aligned}$$

i.e. any two $\langle Au_1,v_1\rangle,\langle Au_2,v_2\rangle$ are uncorrelated. Using this covering set, we have

$$\begin{aligned}\mathbb{E}\left[\|A\|_{\text{op}}\right] \geq & \mathbb{E}\left[\sup_{u\in\mathcal{U},v\in\mathcal{V}}\langle Au,v\rangle\right]\\ \overset{(*)}{\gtrsim} & \sqrt{\log|\mathcal{U}|+\log|\mathcal{V}|}\\ \gtrsim & \sqrt{m}+\sqrt{n}\end{aligned}$$

where $(*)$ by Martin J. Wainwright's book, page 53, Exercise 2.11 (Upper and lower bounds for Gaussian maxima).

As a result, we have

$$\mathbb{E}\left[\|A\|_{\text{op}}\right] \asymp \sqrt{m}+\sqrt{n}$$

Exercise 4 Rademacher Complexity Leads to Suboptimal Bounds

4.(a)

We have by CLT:

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N(0,I)$$

and we have

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|_2^2\right] = \frac{d}{n}$$

4.(b)

We have the hypothesis space $\Theta = \mathbb{R}^d$. We consider the following:

1. Bound covering number: We consider the covering of $\mathbb{B}_n(\delta,\Theta)$ under norm $\frac{1}{n}\sum_{i=1}^n (X_i - \cdot)^2$:

$$\log N(t, \mathbb{B}_n(\delta,\Theta), \|\cdot\|_2) \leq d\log(1+\frac{2\delta}{t}).$$

2. We determine some δ satisfying the following:

$$\begin{aligned}
&\frac{16}{\sqrt{n}}\int_0^\delta \sqrt{\log N(t, \mathbb{B}_n(\delta,\Theta), \|\cdot\|_2)}\,\mathrm{d}t \leq \frac{\delta^2}{4}\\
\Leftarrow &\frac{1}{\sqrt{n}}\int_0^\delta \sqrt{d\log(1+\frac{2t}{\delta})}\,\mathrm{d}t \lesssim \delta^2\\
\Leftarrow &\frac{\sqrt{d}}{\sqrt{n}}\delta \lesssim \delta^2\\
\Leftarrow &\delta \gtrsim \sqrt{\frac{d}{n}}
\end{aligned}$$

3. By Martin J. Wainwright's book, page 426, Corollary 13.7, we have

$$\mathbb{E}\left[\left\|\hat{\theta}-\theta\right\|_n^2\right] \lesssim \frac{d}{n}$$

So we see that we have the same optimal rate of convergence, but the constant is undetermined.

Exercise 5 Curse of Dimensionality

We consider the following segment to $[-B,+B]$ and partition of $[-R,+R]^d$:

- $[-B,+B]$ into segments of length ε, which gives $\lceil\frac{2B}{\varepsilon}\rceil$ segments (with trivial edges omitted). The set is denoted as $\mathcal{S} = \{S_1,\ldots,S_{\lceil\frac{2B}{\varepsilon}\rceil}\}$.
- $[-R,+R]^d$ into cubes of side length $\varepsilon/2\rho$, which gives $\lceil\frac{4\rho(R+\varepsilon)}{\varepsilon}\rceil^d$ cubes (with trivial edges omitted). So we haved a lattice $\mathcal{C}$ of side length ε/ρ.

For any function f in function class $\mathcal{F} : [-R, +R]^d \to [-B, +B]$, there exists some $S_f \in \mathcal{S}$ s.t. $|f(0) - S_f| \leq \varepsilon/2$. Next, for any $x \in \mathcal{C}$, we do the following:

1. Find the $S_x \in \mathcal{S}$ s.t. $|f(x) - S_x| \leq \varepsilon/2$.
2. For any $x' \in x + \varepsilon/2\rho \cdot [-1, 1]^d$, we have

$$\left|f(x') - S_x\right| \leq \left|f(x') - f(x)\right| + |f(x) - S_x| \leq \varepsilon/2 + \varepsilon/2 = \varepsilon$$

So at least one of $S_x, S_x + \varepsilon, S_x - \varepsilon$ would satisfy:

$$\forall x' \in x + \varepsilon/2\rho \cdot [-1, 1]^d, \quad \exists S_{x'} \in \{S_x, S_x + \varepsilon, S_x - \varepsilon\} \quad \left|f(x') - S_{x'}\right| \leq \varepsilon$$

By doing so iteratively, we can determine that there are $3^{|\mathcal{C}|}$ possible functions in the covering set (after $f(0)$ is given).

So in total, we have the size of the covering set begin:

$$\begin{aligned}
\log \mathcal{N}(\mathcal{F}, \varepsilon, \|\cdot\|_\infty) \leq & \log \left[\lceil \frac{2B}{\varepsilon} \rceil \times 3^{|\mathcal{C}|} \vee 1 \right] \\
\leq & \log \left[\lceil \frac{2B}{\varepsilon} \rceil \times 3^{\lceil \frac{4\rho(R+\varepsilon)}{\varepsilon} \rceil^d} \vee 1 \right] \\
= & 0 \vee \left[\lceil \frac{4\rho(R+\varepsilon)}{\varepsilon} \rceil^d \log 3 + \log \lceil \frac{2B}{\varepsilon} \rceil \right]
\end{aligned}$$

(which is a stronger bound than required in the homework question).

IEMS 402 Statistical Learning - 2025 Winter
HW7

Tuorui Peng[1]

Exercise 1 Dvoretzky-Kiefer-Wolfowitz inequality via Uniform Bounds

1.(a)

For each given P, consider the $\{\varepsilon, 2\varepsilon, \dots, \lfloor 1/\varepsilon \rfloor \varepsilon\}$ quantiles of P, denoted by $q_1, q_2, \dots, q_m$, where $m = \lfloor 1/\varepsilon \rfloor$. With this notation, we have

$$\{\mathbf{1}\{x \le q_i\}\}_{i=1}^m$$

being a ε covering of $\mathcal{F} = \{\mathbf{1}\{x \le t\}\}_{t\in\mathbb{R}}$. Thus we have

$$\sup_P \log N(\mathcal{F}, L_2(P), \varepsilon) \le \log m \lesssim \log(1 + \frac{1}{\varepsilon})$$

1.(b)

When upgrading to n points, the above covering number bound becomes

$$\log N(\mathcal{F}^n, L_2(P), \varepsilon) \lesssim n \log(1 + \frac{1}{\varepsilon})$$

With the covering number bound, we apply the following:

$$\begin{aligned}
R_n(\mathcal{F}) =& \mathbb{E}\left[\sup_{t\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n \mathbf{1}\{X_i \le t\} - P(t) \right| \right] \\
\le& \frac{2}{n}\mathbb{E}\left[\sup_{F\in\mathcal{F}} \left|\langle \varepsilon, F_{X_1^n}(t)\rangle\right|\right] \\
\overset{\text{Dudley}}{\lesssim}& \frac{1}{n}\int_0^1 \sqrt{\log N(\mathcal{F}^n, L_2(P), u)}\, \mathrm{d}u \\
\lesssim& \int_0^1 \sqrt{n \log(1 + \frac{1}{u})}\, \mathrm{d}u \\
\lesssim& \frac{1}{\sqrt{n}}
\end{aligned}$$

1.(c)

Note that $P_n(\,\cdot\,)$ is sub-Gaussian, we can apply the following:

$$\left| \sup_{t\in\mathbb{R}} |P_n(X \le t) - P(X \le t)| - \mathbb{E}\left[\sup_{t\in\mathbb{R}} |P_n(X \le t) - P(X \le t)|\right] \right| \le \varepsilon \quad \text{w.p. at least } 1 - 2\exp\left(-\frac{2n\varepsilon^2}{C^2}\right)$$

i.e. with the above bound on expectation, we have

$$\sup_{t\in\mathbb{R}} |P_n(X \le t) - P(X \le t)| \ge \frac{C}{\sqrt{n}} + \varepsilon \quad \text{w.p. at most } 2\exp\left(-cn\varepsilon^2\right)$$

[1] TuoruiPeng2028@u.northwestern.edu

Exercise 2

- Assume we have a set that form a 2ε packing $\mathcal{S}(2\varepsilon)$ and we take any two points $m_1, m_2 \in \mathcal{S}(2\varepsilon)$ and any $h \in \mathbb{B}(h_i, \varepsilon)$, then

$$\|h - h_i\| \geq \|h_i - h_j\| - \|h - h_i\| > 2\varepsilon - \varepsilon = \varepsilon \quad \forall h \in \mathbb{B}(h_i, \varepsilon)$$

  which means that $\mathcal{S}(2\varepsilon)\backslash\{h_i\}$ is not a ε-covering. By ranging over all such $\mathcal{S}$ we would eventually have some set satisfying the maximal packing, however in any of the setting we see that $|\mathcal{S}| \leq |\mathcal{N}(\varepsilon)|$. And we can conclude that

$$M(2\varepsilon) \leq N(\varepsilon)$$

- Consider the maximal packing $\mathcal{M}\varepsilon$. For any $h \in \mathcal{H}$ we have

$$\exists h_i \in \mathcal{M},\ st.\ \|h - h_i\| \leq \varepsilon$$

  (otherwise we have $\mathcal{M} \cup \{h\}$ as a larger packing). Thus we directly see that this set forms a ε-covering, and we have

$$N(\varepsilon) \leq M(\varepsilon)$$

To summarize, we have

$$M(2\varepsilon) \leq N(\varepsilon) \leq M(\varepsilon)$$

Exercise 3

- For the minimal covering set $\mathcal{N}(\Theta)$ we have:

$$\text{vol}(\Theta) \leq |\mathcal{N}(\Theta)| \cdot \text{vol}(B(\varepsilon)) \Rightarrow |\mathcal{N}(\Theta)| \geq \frac{\text{vol}(\Theta)}{\text{vol}(B(\varepsilon))}$$

- For the maximal packing set $\mathcal{M}(\Theta)$ we have:

$$\text{vol}(\Theta + B(\varepsilon/2)) \geq |\mathcal{M}(\Theta)| \cdot \text{vol}(B(\varepsilon)) \Rightarrow |\mathcal{M}(\Theta)| \leq \frac{\text{vol}(\Theta + B(\varepsilon/2))}{\text{vol}(B(\varepsilon))}$$

So we have

$$\frac{\text{vol}(\Theta)}{\text{vol}(B(\varepsilon))} \leq |\mathcal{N}(\Theta)| \leq |\mathcal{M}(\Theta)| \leq \frac{\text{vol}(\Theta + B(\varepsilon/2))}{\text{vol}(B(\varepsilon))}$$

Exercise 4 Covering Number of Sobolev Ellipsoid

4.(a)

In this setting, we have for any pair $\theta^1 \in \tilde{\mathcal{E}}$ and any $\theta \in \mathcal{E}$:

$$\begin{aligned}\left\|\theta-\theta^1\right\|_2^2 =& \sum_{i=1}^{t}(\theta_i-\theta_i^1)^2+\sum_{i=t+1}^{d}\theta_i^2 \\ \leq& \delta^2+\sum_{i=t+1}^{d}\theta_i^2 \leq \qquad\qquad \delta^2+\delta^2\sum_{i=t+1}^{d}\frac{1}{\mu_j} \\ \leq& 2\delta^2\end{aligned}$$

i.e. by choosing $t : \mu_t \leq \delta^2$, we have the δ-covering of $\tilde{\mathcal{E}}$ satisfying $\sqrt{2}\delta$-covering of $\mathcal{E}$.

4.(b)

As stated above, it suffices to bound the truncated ellipsoid $\tilde{\mathcal{E}}$, which has "finite" dimension t and we know that the covering number of $\tilde{\mathcal{E}}$ is bounded by

$$N(\tilde{\mathcal{E}},\delta) \lesssim \left(\frac{1}{\delta}\right)^t$$

in which t satisfies $\mu_t = t^{-2\alpha} \asymp \delta^2 \Rightarrow t \asymp \delta^{-1/\alpha}$.

Adn we have

$$N(\tilde{\mathcal{E}},\delta) \leq N(\tilde{\mathcal{E}},\delta/\sqrt{2}) \lesssim \left(\frac{\sqrt{2}}{\delta}\right)^{\delta^{-1/\alpha}}$$

i.e.

$$\log N(\tilde{\mathcal{E}},\delta) \lesssim \left(\frac{1}{\delta}\right)^{1/\alpha}\log\frac{1}{\delta}$$

IEMS 402 Statistical Learning - 2025 Winter
HW8

Tuorui Peng[1]

Exercise 1 Hilbert Embedding of Probability

1.(a)

Consider the functional $L : f \mapsto \mathbb{E}[f(X)]$ which is a bounded linear functional. By Riesz Representation Theorem, there exists a unique $h_L \in \mathcal{H}$ such that

$$\forall f \in \mathcal{H} : \langle h_L, f\rangle = L(f) = \mathbb{E}_P[f(X)] = \mathbb{E}_P[\langle \varphi(X), f\rangle] = \langle \mathbb{E}_P[\varphi(X)], f\rangle$$

i.e. such $\mathcal{H} \ni h_L = \mathbb{E}_P[\varphi(X)]$.

1.(b)

We prove the contrapositive. Suppose $\mathbb{E}_P[\varphi(X)] = \mathbb{E}_Q[\varphi(X)]$, then we consider the following setting: $\forall \varepsilon > 0$, $\forall f \in \mathcal{X}$, $\exists h_{f,\varepsilon} \in \mathcal{H}$ s.t. $\|f - h_{f,\varepsilon}\|_\infty < \varepsilon$, and we have

$$\begin{aligned}
\mathbb{E}_P[h_{f,\varepsilon}(X)] =& \mathbb{E}_P[\langle h_{f,\varepsilon}, \varphi(X)\rangle] \\
=& \langle \mathbb{E}_P[\varphi(X)], h_{f,\varepsilon}\rangle \\
\mathbb{E}_Q[h_{f,\varepsilon}(X)] =& \mathbb{E}_Q[\langle h_{f,\varepsilon}, \varphi(X)\rangle] \\
=& \langle \mathbb{E}_Q[\varphi(X)], h_{f,\varepsilon}\rangle
\end{aligned}$$

So we have

$$\begin{aligned}
|\mathbb{E}_P[f(X)] - \mathbb{E}_Q[f(X)]| \leq& 2\varepsilon + |\mathbb{E}_P[h_{f,\varepsilon}(X)] - \mathbb{E}_Q[h_{f,\varepsilon}(X)]| \\
=& 2\varepsilon + |\langle \mathbb{E}_P[\varphi(X)] - \mathbb{E}_Q[\varphi(X)], h_{f,\varepsilon}\rangle| \\
=& 2\varepsilon
\end{aligned}$$

Note that the above statement is true $\forall \varepsilon > 0$, $\forall f \in \mathcal{C}$, thus proves the contrapositive that $P \overset{\mathrm{d}}{=} Q$ must hold. And we have if $P \overset{\mathrm{d}}{\neq} Q$, then $\mathbb{E}_P[\varphi(X)] \neq \mathbb{E}_Q[\varphi(X)]$.

1.(c)

We have for right hand side:

$$\begin{aligned}
\text{R.H.S.} =& \sqrt{\mathbb{E}[k(X, X')] + \mathbb{E}[k(Z, Z')] - 2\mathbb{E}[k(X, Z)]} \\
=& \sqrt{\langle \varphi(X), \varphi(X')\rangle + \langle \varphi(Z), \varphi(Z')\rangle - 2\langle \varphi(X), \varphi(Z)\rangle} \\
=& \sqrt{\langle \varphi(X) - \varphi(X'), \varphi(Z) - \varphi(Z')\rangle} \\
:=& E
\end{aligned}$$

[1] TuoruiPeng2028@u.northwestern.edu

For left hand side:

$$\sup_{f\in\mathcal{H},\|f\|\leq 1}\left|\mathbb{E}_P\left[f(X)\right]-\mathbb{E}_Q\left[f(X)\right]\right|=\sup_{f\in\mathcal{H},\|f\|\leq 1}\left|\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],f\right\rangle\right|$$

which reach maximum when $f\propto\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]:=\alpha\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]$, in which α is taken to make $\|f\|=1$. Thus we have

$$1=\alpha^2\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]\right\rangle=\alpha^2E^2\Rightarrow\alpha=\frac{1}{E}$$

Substitute back to the left hand side, we have

$$\begin{aligned}\text{L.H.S.}=&\sup_{f\in\mathcal{H},\|f\|\leq 1}\left|\mathbb{E}_P\left[f(X)\right]-\mathbb{E}_Q\left[f(X)\right]\right|\\=&\left|\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],\frac{1}{E}\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]\right\rangle\right|\\=&\frac{1}{E}\left\langle\mathbb{E}\left[\varphi(X)-\varphi(Z)\right],\mathbb{E}\left[\varphi(X)-\varphi(Z)\right]\right\rangle\\=&\frac{1}{E}E=E=\text{R.H.S.}\end{aligned}$$

Exercise 2 Example of Kernel

2.(a)

We can verify the condition for k_{norm} to be a kernel function by checking the positive semi-definiteness and symmetry easily:

$$\begin{aligned}k_{\text{norm}}(x,z)=&k_{\text{norm}}(z,x)\\\forall x_1^m,\,\alpha_1^n,\,\sum_{i,j=1}^n k_{\text{norm}}(x_i,x_j)\alpha_i\alpha_j=&\sum_{i,j=1}^n k(x_i,z_i)\frac{\alpha_i}{\sqrt{k}(x_i,x_i)}\frac{\alpha_j}{\sqrt{k(x_j,x_j)}}\geq 0\end{aligned}$$

2.(b)

We prove the reproducing property by checking the following:

$$\begin{aligned}\forall f,\,\forall x\,:\,\langle k(x,\,\cdot\,),f\rangle=&\int_0^1 k'(x,z)f'(z)\,\mathrm{d}z\\=&\int_0^1\mathbf{1}_{[0,x]}(z)\mathbf{1}_{[0,x]}(z)f'(z)\,\mathrm{d}z\\=&\int_0^x f'(z)\,\mathrm{d}z\\=&f(x)\end{aligned}$$

And we can easily verify the symmetry and positive semi-definiteness of $k(\,\cdot\,,\,\cdot\,) = \cdot \wedge \cdot$ as follows:

$$\begin{aligned} k(x,z) =& x \wedge z = z \wedge x = k(z,x) \\ \forall g: \int_0^1 \int_0^1 g(x)g(z)k(x,z)\,\mathrm{d}x\,\mathrm{d}z =& \int_0^1 \int_0^1 g(x)g(z)\left\langle \mathbf{1}_{[0,x]}, \mathbf{1}_{[0,z]} \right\rangle \mathrm{d}x\,\mathrm{d}z \\ =& \int_0^1 \int_0^1 \left\langle g(x)\mathbf{1}_{[0,x]}, g(z)\mathbf{1}_{[0,z]} \right\rangle \mathrm{d}x\,\mathrm{d}z \\ =& \left\| \int_0^1 g(x)\mathbf{1}_{[0,x]}\,\mathrm{d}x \right\|^2 \geq 0. \end{aligned}$$

2.(c)

WLOG take $f^{(i)}(0) = 0$ for all $i \leq k-1$. So that $\langle f, g\rangle = \int_0^1 f^{(k)}(x)g^{(k)}(x)\,\mathrm{d}x$.

Using similar integration by parts idea, we should have: for each given x, the function $k(x,\,\cdot\,)$ satisfies

$$\begin{aligned} g(x) =& \langle k(x,\,\cdot\,), g\rangle \\ =& \int_0^1 k^{(k)}(x,z)g^{(k)}(z)\,\mathrm{d}z \end{aligned}$$

By the Taylor expansion of g at 0 with integraion remainders, i.e.

$$g(x) = \int_0^x \frac{g^{(k)}(z)}{(k-1)!}(x-z)^{k-1}\,\mathrm{d}z$$

we have

$$k^{(k)}(x,z) = \frac{(x-z)^{k-1}}{(k-1)!}\mathbf{1}_{[0,x]}(z) = \frac{(x-z)_+^{k-1}}{(k-1)!}$$

and thus

$$\begin{aligned} k(x,z) =& \langle k(x,\,\cdot\,), k(z,\,\cdot\,)\rangle \\ =& \int_0^1 k^{(k)}(x,u)k^{(k)}(z,u)\,\mathrm{d}u \\ =& \int_0^1 \frac{(x-u)_+^{k-1}}{(k-1)!}\frac{(z-u)_+^{k-1}}{(k-1)!}\,\mathrm{d}u \end{aligned}$$

Exercise 3 φ-divergence DRO and Variance Regularization

Optimzation problem is formalized as:

$$\sup_{P\in\mathcal{P}_n} \mathbb{E}_P\left[\ell(\theta, X)\right], \quad s.t.\, D_\varphi(P\|\hat{P}_n) \leq \frac{\rho}{n}$$

Since the empirical distribution $\hat{P}_n$ is a Dirac measure, the optimizer would also be a Dirac measure supported on $\mathrm{supp}(X_1^n)$. We denoted this PMF as:

$$P:\ X = X_i, \quad w.p.\, p_i$$

So the optimization problem can be reformulated as:

$$\sup_{p_i\geq 0,\sum_{i=1}^n p_i=1}\sum_{i=1}^n p_i\ell(\theta,X_i),\quad s.t.\, D_\varphi(P\|\hat{P}_n)\leq\frac{\rho}{n}$$

Lagrangian:

$$\mathcal{L}(\vec{p};\lambda,\mu)=\sum_{i=1}^n p_i\ell(\theta,X_i)+\lambda\left(\sum_{i=1}^n np_i^2-1-\frac{2\rho}{n}\right)+\mu\left(\sum_{i=1}^n p_i-1\right),\quad \lambda\leq 0$$

which is maximized w.r.t. $\vec{p}$ when

$$p_i^*=-\frac{\ell(X_i,\theta)+\mu}{2\lambda n}$$

and gives dual problem:

$$\theta_D(\lambda,\mu)=-\sum_{i=1}^n\frac{(\ell(X_i,\theta)+\mu)^2}{4\lambda n}-\lambda(1+\frac{2\rho}{n})-\mu,\quad \lambda\geq 0$$

Which is minimized w.r.t. λ,μ:

$$0=\begin{cases}\frac{\partial}{\partial\lambda}\theta_D= & \sum_{i=1}^n\frac{(\ell(X_i,\theta)+\mu)^2}{4\lambda^2 n}-(1+\frac{2\rho}{n})\\ \frac{\partial}{\partial\mu}\theta_D= & -\sum_{i=1}^n\frac{(\ell(X_i,\theta)+\mu)}{2\lambda n}-1\end{cases}\Rightarrow\begin{cases}\mu=\sqrt{\frac{var_{\hat{P}_n}(\ell(X,\theta))}{2\rho/n}}-\mathbb{E}_{\hat{P}_n}[\ell(X,\theta)]\\ \lambda=-\frac{1}{2}\sqrt{\frac{var_{\hat{P}_n}(\ell(X,\theta))}{2\rho/n}}\end{cases}$$

and gives (with strong duality):

$$\begin{aligned}R_n(\theta,\mathcal{P}_n)=&\sup_{P\in\mathcal{P}_n}\mathbb{E}_P[\ell(\theta,X)]=\inf_{\lambda\leq 0,\mu}\theta_D(\lambda,\mu)\\ =&\mathbb{E}_{\hat{P}_n}[\ell(X,\theta)]+\sqrt{\frac{2\rho\mathrm{var}_{\hat{P}_n}(\ell(X,\theta))}{n}}\end{aligned}$$

And we revisit the solution p_i^*:

$$p_i^*=-\frac{\ell(X_i,\theta)+\mu}{2\lambda n}=\frac{1}{n}+\frac{\sqrt{2\rho}(\ell(X_i,\theta)-\mathbb{E}_{\hat{P}_n}[\ell(X,\theta)])}{n\sqrt{n}\sqrt{\mathrm{var}_{\hat{P}_n}(\ell(X,\theta))}}$$

could satisfy $p_i^*\geq 0$ if the empirical variance is large enough & ρ is chosen small enough.

Exercise 4 Derive the dual formulation of the Sinkhorn distance

4.(a)

The Lagrangian is given by:

$$\mathcal{L}(\gamma;u,v)=\langle\gamma,C\rangle-\varepsilon H(\gamma)+u'(\gamma\mathbf{1}-a)+v'(\gamma^T\mathbf{1}-b)$$

miimizing w.r.t. γ gives the optimality condition:

$$\frac{\partial\mathcal{L}}{\partial\gamma_{ij}}=C_{ij}+\varepsilon(\log\gamma_{ij}+1)+u_i+v_j=0\Rightarrow\gamma_{ij}=\exp\left(-\frac{C_{ij}+u_i+v_j}{\varepsilon}-1\right)$$

And we have dual problem:

$$\begin{aligned}\inf_{u,v}\mathcal{L}(\gamma;u,v) =& \inf_{u,v}\langle\gamma,C\rangle-\varepsilon H(\gamma)+u'(\gamma\mathbf{1}-a)+v'(\gamma^T\mathbf{1}-b)\\ =& \inf_{u,v} \quad -u'a-v'b-\varepsilon\sum_{i,j}\exp\left(-\frac{C_{ij}+u_i+v_j}{\varepsilon}-1\right)\end{aligned}$$

with re-parametrization $u\mapsto -u-\varepsilon, \quad v\mapsto -v$ and omitting constant, we have dual problem:

$$\inf_{u,v} \quad u'a+v'b-\varepsilon\sum_{i,j}\exp\left(\frac{u_i+v_j-C_{ij}}{\varepsilon}\right)$$

4.(b)

Once the optimal u^* is known, it's left to solve for v^*:

$$\begin{aligned}&\inf_{v} \quad v'b-\varepsilon\sum_{i,j}\exp\left(\frac{u_i^*+v_j-C_{ij}}{\varepsilon}\right)v\\ \Rightarrow& v_j^*=\frac{\varepsilon}{b_j}\sum_i\exp\left(\frac{u_i^*+v_j^*-C_{ij}}{\varepsilon}\right)\Rightarrow v^*, \quad \forall j\end{aligned}$$

which can be solved by fixed point iteration.