Question 1.1: Noise and True Risk (10 Points)

Let $f \in \mathcal{H}$ with a bounded loss function (as per Eq. (1.3)). Prove that

$$|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]| \leq D_{TV}(p, \widetilde{p}), \tag{1.5}$$

where $p, \widetilde{p}$ are the density functions of $\mathcal{D}, \widetilde{\mathcal{D}}$, respectively; and

$$D_{\mathrm{TV}}(p, \widetilde{p}) \stackrel{\text{def}}{=} \iint |p(x, y) - \widetilde{p}(x, y)| \, \mathrm{d}x \, \mathrm{d}y. \tag{1.6}$$

$$\begin{aligned}
|R_D[f] - R_{\tilde{D}}[f]| &= \left| \mathop{E}_{(x,y)\sim D}[\ell(f(x), y)] - \mathop{E}_{(x,y)\sim \tilde{D}}[\ell(f(x), y)] \right| \\
&= \left| \iint \ell(f(x), y) \cdot p(x, y) \, dx dy - \iint \ell(f(x), y) \cdot \tilde{p}(x, y) \, dx dy \right| \\
&= \left| \iint \ell(f(x), y) \cdot p(x, y) - \ell(f(x), y) \cdot \tilde{p}(x, y) \, dx dy \right| \\
&= \left| \iint \ell(f(x), y) \cdot (p(x, y) - \tilde{p}(x, y)) \, dx dy \right|
\end{aligned}$$

$\because \ell(y, y) \in [0, 1]$ $\quad |R_D[f] - R_{\tilde{D}}[f]|$

$$\therefore 0 = |0| \leq \left| \iint \ell(f(x), y) \cdot (p(x, y) - \tilde{p}(x, y)) \, dx dy \leq \left| \iint (p(x, y) - \tilde{p}(x, y)) \, dx dy \right| \right.$$

$\because \forall (x, y) \in D \quad |p(x, y) - \tilde{p}(x, y)| \geq p(x, y) - \tilde{p}(x, y)$

$$\therefore \iint |p(x, y) - \tilde{p}(x, y)| \, dx dy \geq \iint p(x, y) - \tilde{p}(x, y) \, dx dy$$

$\because \forall (x, y) \in D \quad |p(x, y) - \tilde{p}(x, y)| \geq -(p(x, y) - \tilde{p}(x, y))$

$$\therefore \iint |p(x, y) - \tilde{p}(x, y)| \, dx dy \geq \iint -(p(x, y) - \tilde{p}(x, y)) \, dx dy = -\iint (p(x, y) - \tilde{p}(x, y)) \, dx dy$$

$$\therefore -\iint |p(x, y) - \tilde{p}(x, y)| \, dx dy \leq \iint p(x, y) - \tilde{p}(x, y) \, dx dy \leq \iint |p(x, y) - \tilde{p}(x, y)| \, dx dy$$

$$\therefore \left| \iint p(x, y) - \tilde{p}(x, y) \, dx dy \right| \leq \iint |p(x, y) - \tilde{p}(x, y)| \, dx dy = DTV(p, \tilde{p}) \Rightarrow |R_D[f] - R_{\tilde{D}}[f]| \leq DTV(p, \tilde{p})$$

Question 1.2: Uniform Bound for Noisy Sample Gap (15 Points)

Let $\mathcal{H}$ be a finite hypothesis class. Prove that given a finite sample $\mathcal{S} \sim \mathcal{D}^N$, for $\delta > 0$

$$\Pr\left\{ \forall f \in \mathcal{H} : |\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\mathcal{S}}[f]| \leq D_{\mathrm{TV}}(p, \widetilde{p}) + \sqrt{\frac{\ln|\mathcal{H}| + \ln(1/\delta)}{N}} \right\} \geq 1 - \delta. \tag{1.7}$$

$P\left(\forall f \in H : |R_D[f] - \hat{R}_S[f]| \leq t\right) \geq 1 - \delta$, To prove $t = D_{TV}(P, \tilde{P}) + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}}$

$$\because |R_D[f] - \hat{R}_S[f]| = |R_D[f] - R_{\tilde{D}}[f] + R_{\tilde{D}}[f] - \hat{R}_S[f]| \leq |R_D[f] - R_{\tilde{D}}[f]| + |R_{\tilde{D}}[f] - \hat{R}_S[f]|$$

$$\because |R_D[f] - R_{\tilde{D}}[f]| \leq D_{TV}(P, \tilde{P})$$

$$\therefore |R_D[f] - \hat{R}_S[f]| \leq D_{TV}(P, \tilde{P}) + |R_{\tilde{D}}[f] - \hat{R}_S[f]|$$

From the Proposition 1, we know that $P(|R_{\tilde{D}}[f] - \hat{R}_S[f]| \geq \epsilon) \leq 2\exp(-2N\epsilon^2)$

Then, $P(|R_{\tilde{D}}[f] - \hat{R}_S[f]| \leq \epsilon) \geq 1 - 2\exp(-2N\epsilon^2)$

$\therefore 1-2\exp(-2N\epsilon^2) = 1-\delta$

$\therefore \epsilon = \sqrt{\frac{\ln(2/\delta)}{2N}} \leq \sqrt{\frac{\ln 2 + \ln(1/\delta)}{N}}$

From piazza, Josh Nguyen said that we can assume $|H| \geq 2$ and $\delta \in (0,1)$

So $\ln|H| \geq \ln 2$

$\therefore \epsilon \leq \sqrt{\frac{\ln 2 + \ln(1/\delta)}{N}} \leq \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}} \quad \forall f \in H$

$\therefore |R_{\tilde{D}}[f] - \hat{R}_{\tilde{S}}[f]| \leq \epsilon \leq \sqrt{\frac{\ln(H) + \ln(1/\delta)}{N}} \quad \forall f \in H$

So $|R_D[f] - \hat{R}_{\tilde{S}}[f]| \leq D_{TV}(P, \tilde{P}) + |R_{\tilde{D}}[f] - \hat{R}_{\tilde{S}}[f]| \leq D_{TV}(P, \tilde{P}) + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}} \quad \forall f \in H$

So $P\left(\forall f \in H : |R_D[f] - \hat{R}_{\tilde{S}}[f]| \leq D_{TV}(P, \tilde{P}) + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}}\right) \geq 1-\delta$

Question 1.3: Interpreting the Uniform Bound for Noisy Sample Gap (10 Points)

Compare Eq. (1.7) with the bound we found in the lecture in the absence of noisy data, namely, for any $\delta > 0$,

$$\Pr\left\{\forall f \in \mathcal{H} : |\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\mathcal{S}}[f]| \leq \sqrt{\frac{\ln|\mathcal{H}| + \ln(1/\delta)}{N}}\right\} \geq 1 - \delta.$$

What is the effect of noisy data on the generalisation bound and the rate at which it decreases with the number of data points?

Compare the Uniform Bound for the Noisy Sample Gap with the bound found in lecture.
We find that, the noisy data increases the generalisation bound by $D_{TV}(P, \tilde{P})$, which measure the total difference between clean data distribution pdf and the noisy data distribution pdf. if the $\tilde{P} = P$, then Bound for the Noisy Sample Gap equals to the bound introduced in the lecture. If it has more noisy, the bound will get bigger and bigger.

The rate at which it decreases with the number of data points is $\sqrt{\frac{1}{N}}$, because $D_{TV}(P, \tilde{P})$ is constant and the term $\sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}}$ tells me $\sqrt{\frac{c}{N}}$, so the rate is $\sqrt{\frac{1}{N}}$

Question 1.4: Noisy Posterior for Label Noise (5 Points)

Given a clean distribution $\mathcal{D}$ and a noisy version of the distribution $\widetilde{\mathcal{D}}$ under asymmetric label noise with flipping probabilities $\sigma_{-1}, \sigma_{+1} \in [0,1]$, show that for all $y \in \mathcal{Y}$

$$\Pr(\widetilde{Y} = y \mid X) = \sigma_{-y} \cdot (1 - \Pr(Y = y \mid X)) + (1 - \sigma_y) \cdot \Pr(Y = y \mid X). \tag{1.8}$$

From (LN3), we know $\widetilde{Y} \perp\!\!\!\perp X \mid Y$, so we know $P(\widetilde{Y}, X \mid Y) = P(\widetilde{Y} \mid Y)\, P(X \mid Y)$
and $P(\widetilde{Y} \mid X, Y) = P(\widetilde{Y} \mid Y)$ – (LN3.1)
for all $y \in \mathcal{Y} = \{-1, +1\}$

$$P(\widetilde{Y}=y \mid X) = \sum_y P(\widetilde{Y}=y, Y=y \mid X) = P(\widetilde{Y}=y, Y=y \mid X) + P(\widetilde{Y}=y, Y=-y \mid X)$$

$$= P(\widetilde{Y}=y \mid Y=y, X) \cdot P(Y=y \mid X) + P(\widetilde{Y}=y \mid Y=-y, X) \cdot P(Y=-y \mid X)$$

$$= P(\widetilde{Y}=y \mid Y=y) \cdot P(Y=y \mid X) + P(\widetilde{Y}=y \mid Y=-y) \cdot P(Y=-y \mid X) \quad (LN3.1)$$

$$= (1-\sigma_y) \cdot P(Y=y \mid X) + \sigma_{-y} \cdot P(Y=-y \mid X) \quad (LN1)$$

∵ We know Y only can have two values, namely, -1 or $+1$.

∴ $P(Y=y \mid X) + P(Y=-y \mid X) = 1 \equiv P(Y=-y \mid X) = 1 - P(Y=y \mid X)$

∴ $P(\widetilde{Y}=y \mid X) = (1-\sigma_y) \cdot P(Y=y \mid X) + \sigma_{-y}\,(1 - P(Y=y \mid X))$

Question 1.5: Uniform Bound for Symmetric Label Noise (10 Points)

Suppose that a noisy distribution $\widetilde{\mathcal{D}}$ is given by symmetric label noise, with flipping $\sigma \in [0,1]$, applied to clean distribution $\mathcal{D}$. Prove that given a finite sample $\mathcal{S} \sim \mathcal{D}^N$, for $\delta > 0$

$$\Pr\left\{\forall f \in \mathcal{H} : |\mathcal{R}_{\mathcal{D}}[f] - \widehat{\mathcal{R}}_{\widetilde{\mathcal{S}}}[f]| \le 2\sigma + \sqrt{\frac{\ln|\mathcal{H}| + \ln(1/\delta)}{N}}\right\} \ge 1 - \delta. \tag{1.9}$$

Hint: Start by bounding $|\mathcal{R}_{\mathcal{D}}[f] - \mathcal{R}_{\widetilde{\mathcal{D}}}[f]|$ like in Question 1.1. Then reuse the technique in Question 1.2.

First, try to bound $|R_D[f] - R_{\tilde{D}}[f]| =$

$$= \left| \iint \ell(f(x), y) \cdot p(x, y)\, dx\, dy - \iint \ell(f(\tilde{x}), \tilde{y}) \cdot p(\tilde{x}, \tilde{y})\, d\tilde{x}\, d\tilde{y} \right| =$$

(supplementary proof of $P(\tilde{Y} \mid \tilde{X}) = P(\tilde{Y} \mid X)$)
LN2 says that $p(\tilde{x} \mid \tilde{y}) = p(x \mid \tilde{y})$
so $P(\tilde{Y} \mid \tilde{X}) = \frac{P(\tilde{x} \mid \tilde{Y}) \cdot P(\tilde{Y})}{P(\tilde{x})} = \frac{P(x \mid \tilde{Y}) \cdot P(\tilde{Y})}{P(x)}$ (LN2)
$= P(\tilde{Y} \mid x)$
∴ $P(\tilde{Y} \mid \tilde{X}) = P(\tilde{Y} \mid X)$

$$= \left| \int \left[\ell(f(x), +1)\, p(x, y=+1) + \ell(f(x), -1)\, p(x, y=-1) \right] dx - \int \left[\ell(f(\tilde{x}), +1) \cdot p(\tilde{x}, \tilde{y}=+1) + \ell(f(\tilde{x}), -1) \cdot p(\tilde{x}, \tilde{y}=-1) \right] d\tilde{x} \right|$$

$$= \left| \int \left[\ell(f(x), +1) \cdot P(y=+1 \mid x) p(x) + \ell(f(x), -1) \cdot P(y=-1 \mid x) p(x) \right] dx - \cdots \right.$$

$$\left. \cdots \int \left[\ell(f(\tilde{x}), +1)\, P(\tilde{y}=+1 \mid x) p(x) + \ell(f(x), -1)\, P(\tilde{y}=-1 \mid x) p(x) \right] dx \right|$$

(LN2 and supplementary proof)

$$= \left| \int p(x) \cdot \left[\ell(f(x), +1) \cdot p(y=+1|x) + \ell(f(x), -1) \cdot p(y=-1|x) \right] dx - \cdots \right.$$

$$\left. \cdots \int p(x) \left[\ell(f(x), +1) \left[6 + (1-2\sigma) p(Y=+1|x) \right] + \ell(f(x), -1) \left[\sigma + (1-2\sigma) \cdot p(Y=-1|x) \right] \right] dx \right|$$

(E.g. 1.8 and $6-1-6+1=6$)

$$= \left| \int p(x) \left[\ell(f(x), +1) \left(2\sigma p(Y=+1|x) - \sigma \right) + \ell(f(x), -1) \left(2\sigma p(Y=-1|x) - \sigma \right) \right] dx \right.$$

$\because 2\sigma p(Y=+1|x) - \sigma \le 2\sigma p(Y=+1|x)$ (because $0 \le \sigma \le 1$ and $2\sigma p(Y=+1|x) \ge 0$)

$\therefore \left| \int p(x) \left[\ell(f(x), +1) \left(2\sigma p(Y=+1|x) - \sigma \right) + \ell(f(x), -1) \left(2\sigma p(Y=-1|x) - \sigma \right) \right] dx \right. \le$

$$\left| \int p(x) \left[\ell(f(x), +1) \cdot 2\sigma \cdot p(Y=+1|x) + \ell(f(x), -1) 2\sigma \cdot p(Y=-1|x) \right] dx \right.$$

$\because \ell(y, y') \in [0, 1]$

$\therefore 0 \le \ell(f(x), +1) \cdot 2\sigma \cdot p(Y=+1|x) \le 2\sigma \cdot p(Y=+1|x)$ and $0 \le \ell(f(x), -1) \cdot 2\sigma \cdot p(Y=-1|x) \le 2\sigma \cdot p(Y=-1|x)$

$\therefore \left| \int p(x) \left[\ell(f(x), +1) \cdot 2\sigma \cdot p(Y=+1|x) + \ell(f(x), -1) 2\sigma \cdot p(Y=-1|x) \right] dx \right. \le \left| \int p(x) [2\sigma p(Y=+1|x) + 2\sigma p(Y=-1|x)] dx \right.$

$$= \left| 2\sigma \int p(x) \cdot (p(Y=+1|x) + p(Y=-1|x)) dx \right| = |2\sigma| = 2\sigma \quad \left(\sum_y^Y p(Y=y|x) = 1, \int p(x) dx = 1 \right)$$

$\therefore \left| R_D[f] - R_B[f] \right| \le 2\sigma$

$P\left(\forall f \in H : \left| R_D[f] - \hat{R}_3[f] \right| \le t \right) \ge 1 - \delta$, To prove $t = 2\sigma + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}}$

$\because \left| R_D[f] - \hat{R}_3[f] \right| = \left| R_D[f] - R_B[f] + R_B[f] - \hat{R}_3[f] \right| \le \left| R_D[f] - R_B[f] \right| + \left| R_B[f] - \hat{R}_3[f] \right|$

$\because \left| R_D[f] - R_B[f] \right| \le 2\sigma$

$\therefore \left| R_D[f] - \hat{R}_3[f] \right| \le 2\sigma + \left| R_B[f] - \hat{R}_3[f] \right|$

From the Proposition 1, we know that $P(|R_B[f] - \hat{R}_3[f]| \ge \epsilon) \le 2\exp(-2N\epsilon^2)$

Then $\overset{\forall f \in H}{\left| R_B[f] - \hat{R}_3[f] \right|} \le \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}}$ (proved in Q1.2)

So $\left| R_D[f] - \hat{R}_3[f] \right| \le 2\sigma + \left| R_B[f] - \hat{R}_3[f] \right| \le 2\sigma + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}} \quad \forall f \in H$

So $P\left(\forall f \in H : \left| R_D[f] - \hat{R}_3[f] \right| \le 2\sigma + \sqrt{\frac{\ln|H| + \ln(1/\delta)}{N}} \right) \ge 1 - \delta$

Question 2.3: Expected Improvement (6 Points)

Derive mathematically the result of the Expected Improvement (EI) acquisition function in Equation 2.13. That is show that,

$$\mathbb{E}\left[\mathrm{I}(\mathbf{x})\right] = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0, \\ 0 & \text{if } \sigma(\mathbf{x}) = 0. \end{cases}$$

$f(x) \geq f(x^+) - \varepsilon$

where Z is defined in Equation 2.14.

① if $\sigma(x) > 0$

$f(x) \sim N(\mu(x), \sigma^2(x)|x)$

$I(x) = \max\{0, f(x) - f(x^+) - \varepsilon\}$

$EI(x) = E_{f(x)\sim N(\mu(x),\sigma^2(x)|x)}[I(x)]$

let $Z_f = \frac{f(x)-\mu(x)}{\sigma(x)}$, so $Z_f \sim N(0,1)$, $Z^+ = \frac{f(x^+)+\varepsilon-\mu(x)}{\sigma(x)}$, $Z = -Z^+ = \frac{\mu(x)-f(x^+)-\varepsilon}{\sigma(x)}$

$EI(x) = E_{Z_f\sim N(0,1)}[I(x)]$

$EI(x) = \int_{-\infty}^{+\infty} I(x)\,\phi(z_f)\,dz_f$

$= \int_{Z^+}^{+\infty} [z_f\sigma(x) + \mu(x) - f(x^+) - \varepsilon]\,\phi(z_f)\,dz_f + 0 \quad (f(x) = z_f\sigma(x) + \mu(x))$

$= \int_{Z^+}^{+\infty} [\mu(x) - f(x^+) - \varepsilon]\,\phi(z_f)\,dz_f - \sigma(x)\cdot\int_{Z^+}^{+\infty} z_f\cdot\phi(z_f)\cdot dz_f$

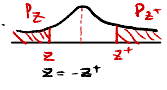
$= [\mu(x) - f(x^+) - \varepsilon]\cdot\int_{Z^+}^{+\infty}\phi(z_f)\,dz_f - \sigma(x)\cdot\int_{Z^+}^{+\infty} z_f\cdot\phi(z_f)\cdot dz_f$

$= [\mu(x) - f(x^+) - \varepsilon]\cdot(1 - \Phi(Z^+)) - \sigma(x)\cdot\int_{Z^+}^{+\infty} z_f\cdot\phi(z_f)\cdot dz_f \quad (1-\Phi(Z^+) = \Phi(Z))$

$= [\mu(x) - f(x^+) - \varepsilon]\cdot\Phi(Z) - \sigma(x)\cdot\int_{Z^+}^{+\infty} z_f\cdot\phi(z_f)\cdot dz_f$

$\because \phi(z_f) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{z_f^2}{2}),\ \frac{d}{dz_f}\exp(-\frac{z_f^2}{2}) = -z_f\cdot\exp(-\frac{z_f^2}{2})$

$= [\mu(x) - f(x^+) - \varepsilon]\cdot\Phi(Z) - \sigma(x)\cdot\int_{Z^+}^{+\infty} z_f\cdot\frac{1}{\sqrt{2\pi}}\exp(-\frac{z_f^2}{2})\,dz_f$

$= [\mu(x) - f(x^+) - \varepsilon]\cdot\Phi(Z) + \frac{\sigma(x)}{\sqrt{2\pi}}\cdot\int_{-\infty}^{Z} -z_f\cdot\exp(-\frac{z_f^2}{2})\,dz_f \quad (\because$  $P_Z = P_{Z^+})$

$= [\mu(x) - f(x^+) - \varepsilon]\cdot\Phi(Z) + \frac{\sigma(x)}{\sqrt{2\pi}}\cdot[\exp(-\frac{Z^2}{2}) - 0]$

$= [\mu(x) - f(x^+) - \varepsilon]\cdot\Phi(Z) + \sigma(x)\cdot\phi(Z)$

② $\sigma(x) = 0 \Rightarrow Z = 0 \Rightarrow 0 = Z^+ = f(x^+) + \varepsilon - \mu(x) \Rightarrow Z = 0 = \mu(x) - f(x^+) - \varepsilon$

$EI(x) = E_{Z_f\sim N(0,1)}[I(x)]$

$EI(x) = \int_{-\infty}^{+\infty} I(x)\,\phi(z_f)\,dz_f$

$= \int_{0}^{+\infty} [z_f\sigma(x) + \mu(x) - f(x^+) - \varepsilon]\,\phi(z_f)\,dz_f + 0$

$$= \int_0^{+\infty} [u(x) - f(x^+) - \varepsilon]\, \phi(z_f)\, dz_f - 0 \cdot \int_0^{+\infty} z_f \cdot \phi(z_f) \cdot dz_f$$

$$= [u(x) - f(x^+) - \varepsilon] \cdot \int_0^{+\infty} \phi(z_f)\, dz_f$$

$$= [u(x) - f(x^+) - \varepsilon] \cdot \Phi(0)$$

$$= 0$$

b) After you run the code on `bayesopt_implementation_viewer.ipynb` for the higher dimensional case, comment on the benefit of including a hyperparameter (meaning the parameters of the covariance function θ of the surrogate model) optimisation approach on your Bayesian Optimisation algorithm. (2 Points)

After runing the bayesopt_implementation_viewer.ipynb, we can observe that the surrogate model can simulate the original function better if we using the hyperparameter optimisation approach, As a result, the next point that it will select will have a higher chance that leads to the global optimum value. Therefore, we can use fewer iterations to do the optimisation compared with the Bayesian Optimisation without using the hyperparameters optimisation.

Another benefit is that it can save time as less iteration, less computation of time-consuming black block function f.

Name: Yixi Ruo (NO GROUP)

UID: u6826541

DATE: 07/05/2022