- $\hat{h}_{n,\theta_b}$: trained single model on dataset with initialization θ_b.
- $\{(x_i, \hat{s}(x_i))\}$, artificial dataset, where $\hat{s}(x) = \mathbb{E}_{\theta_b}\left[s_{\theta_b}(x)\right]$, s_{θ_b} is an θ_b-initialized NN.
- φ'_{n,θ_b}: auxillary network trained on $\{(x_i, \hat{s}(x_i))\}$, denote $\varphi_{n,\theta_b}(x) = \varphi'(x) - \hat{s}(x)$.
- $h^* = \hat{h}_{n,\theta_b} - \varphi_{n,\theta_b}$.

In regression:

$$AU = E(Var(y))$$

$$EU = var(E(y))$$

The key here is when generating artificial dataset, we have

$$\hat{s}(x) = E(s_{\theta_b}(x))$$

$$var(s) = var(s_{\theta_b}(x))$$

The single model $\hat{h}_{n,\theta^b}(x)$:

$$\hat{h}_{n,\theta_b}(x) = S_{\theta^b}(x) + \boxed{K(x,\underline{X})^T \left(K(\underline{X},\underline{X}) + \lambda_n n I \right)^{-1}} \left(y - S_{\theta^b}(\underline{X}) \right)$$

$$h^*(x) = \hat{s}(x) + K(x,\underline{X})^T \left(K(\underline{X},\underline{X}) + \lambda_n n I \right)^{-1} \left(y - \hat{s}(\underline{X}) \right)$$

$$\varphi'(x) = s_{\theta^b}(x) + K(x,\underline{X})^T \left(K(\underline{X},\underline{X}) + \lambda_n n I \right)^{-1} \left(\hat{s}(\underline{X}) - s_{\theta^b}(\underline{X}) \right)$$

$$\Rightarrow \quad \hat{h}^* = \hat{h}_{n,\theta^b} - \varphi'(x) + \hat{s}(x).$$

$$\hat{h}^*(x) = \hat{s}(x) + \underbrace{K(x,X)^T \left(K(X,X) + \lambda_n n I\right)^{-1}}_{K} \left(y - \hat{s}(X)\right)$$

$$\begin{aligned}
var(\hat{h}^*) &= var\left(\hat{s}(x) + K(y - \hat{s}(X))\right) \\
&= var(\hat{s}(x)) + var(K(y - \hat{s}(X))) + 2cov(\cdot,\cdot) \\
&= \frac{1}{m} var(s(x)) + \frac{1}{m} K \cdot var(s(X)) \cdot K^T + 2\, cov(\cdot,\cdot),
\end{aligned}$$

$$\hat{s} = \frac{1}{m} \sum_{i=1}^{m} s_{\theta_{b_i}}(x)$$

$$\begin{aligned}
var(\hat{s}) &= \frac{1}{m^2} var\left(\sum s_i(x)\right) \\
&= \frac{1}{m} \cdot var(s)
\end{aligned}$$

$$\mathrm{Var}(s(x)) \approx \frac{1}{m-1}\sum_{k=1}^{m}\left(s_{\theta^k}(x) - \hat{s}(x)\right)^2$$

$$\mathrm{Var}(s(\underline{X})) \approx \frac{1}{m-1}\sum_{k=1}^{m}\left[s_{\theta^k}(\underline{X}) - \hat{s}(\underline{X})\right]\cdot\left[s_{\theta^k}(\underline{X}) - \hat{s}(\underline{X})\right]^T$$

$$\mathrm{Cov}(s(x), s(\underline{X})) \approx \frac{1}{m-1}\sum_{k=1}^{m}\left[s_{\theta^k}(x) - \hat{s}(x)\right]\left[s_{\theta^k}(\underline{X}) - \hat{s}(\underline{X})\right]^T$$

$$\mathrm{var}(\hat{h}^*) = \frac{1}{m}\cdot\Big[\underbrace{\mathrm{var}(s(x))}_{①\ \text{test data}} + \underbrace{\dot{k}\cdot \underbrace{\mathrm{var}(s(\underline{X}))}_{②\ \text{training data}}\,\dot{k}^T - 2k\,\mathrm{cov}(s(x), s(\underline{X}))}_{③}\Big]$$

training data

① dimension reduction

② decomposation.

$\mathrm{cov} = L^T L$

③ Approximation

measure relation between test data x and training $\underline{X}$

Epistemic.

$$\hat{h}^*(x) = \hat{s}(x) + \underbrace{k(x,\underline{X})^T\left(K(\underline{X},\underline{X}) + \lambda_n n I\right)^{-1}}_{\Downarrow\ k.}(\underline{y} - \hat{s}(\underline{X}))$$

$$\left(\hat{h}^*(x) - \hat{s}(x)\right)^{1\times 1} = k^{1\times n}\cdot\left(\underline{y} - \hat{s}(\underline{X})\right)^{n\times 1} \to t$$

$$\Downarrow$$

$$h$$

$$\Rightarrow \quad h = k \cdot t$$

$$\Rightarrow \quad h\cdot t^T = k\cdot t\cdot t^T$$

$$k = h\cdot t^T (t\cdot t^T)^{-1}$$

$$② = k \cdot var(s(\underline{X})) \cdot k$$

$$= k(x, \underline{X})^T (k(\underline{X}, \underline{X}) + \lambda n n \underline{I})^{-1} \cdot var(s(\underline{X})) \cdot \left[(k(\underline{X}, \underline{X}) + \lambda n n \underline{I})^{-1} \right]^T \cdot k(x, \underline{X})$$

$$\mathrm{var}(\hat{h}^*) = \frac{1}{m} \cdot \Big[$$

$$① \ \mathrm{var}(s(x)) + k \cdot \mathrm{var}(s(\mathbf{X}))\, k^T - 2k\, \mathrm{cov}(s(x), s(\mathbf{X})) \Big]$$

When training data is sufficient $\to \infty$:

① $\mathrm{var}(s(x))$, not change

② $k\, \mathrm{var}(s(\mathbf{X}))\, k^T$, $k(x, \mathbf{X})^T (k(\mathbf{X}, \mathbf{X}) + \lambda n I)^{-1} \cdot \mathrm{var}(s(\mathbf{X})) \cdot ((k(\mathbf{X}, \mathbf{X}) + \lambda n I)^{-1})^T \cdot k(x, \mathbf{X})$

③ $-2k\, \mathrm{cov}(s(x), s(\mathbf{X}))$, $-2k(x, \mathbf{X}) \cdot (k(\mathbf{X}, \mathbf{X}) + \lambda n I)^{-1}\, \mathrm{cov}(s(x), s(\mathbf{X}))$

---

② + ③

$$= k \cdot \mathrm{var}(s(\mathbf{X})) \cdot k^T$$

$$- 2k \cdot \mathrm{cov}(s(x), s(\mathbf{X}))$$

For $K = k(x, \underline{X})\,(k(\underline{X},\underline{X}) + \lambda_n n I)^{-1}$. $\sim k(x, \underline{X}) \cdot O(\frac{1}{n})$

$\lim_{n\to\infty} \frac{1}{n} k(\underline{X},\underline{X}) \to K$

$$(k(\underline{X},\underline{X}) + \lambda_n n I)^{-1}$$
$$= (n(\tfrac{1}{n}k(\underline{X},\underline{X}) + \lambda_n I))^{-1}.$$
$$= \tfrac{1}{n}(\tfrac{1}{n}k(\underline{X},\underline{X}) + \lambda_n I)^{-1} \sim O(\tfrac{1}{n})$$

if x is a point in $\underline{X}$, then

$$k(x, \underline{X}) \approx k(x_i, \underline{X}) \sim O(1).$$

Then $k(x_i, \underline{X}) \cdot (k(\underline{X},\underline{X}) + \lambda_n n I)^{-1}$

$\sim O(\frac{1}{n})$

worst case: $O(\frac{1}{\sqrt{n}})$

So

$$\frac{②+③}{var(s(x))\, var(s(\underline{X}))} = O\left(\frac{1}{\sqrt{n}}\right) \cdot O\left(\frac{1}{\sqrt{n}}\right) \cdot 1 \cdot \frac{1}{var(s(x))} - 2\, O\left(\frac{1}{\sqrt{n}}\right)$$

Now, only thing to do is:

$$var(s(\underline{X})) \cdot O\left(\frac{1}{\sqrt{n}}\right)$$

if $\rightarrow 0$ ✓