Maths of Machine Learning Summary

CHAPTER ONE

Statistical Learning Theory

① **Classification & Regression:** $(X,Y) \in \mathcal{X} \times \mathcal{Y}$ (joint prob measure P_0)

- Goal: learn $h: \mathcal{X} \to \mathcal{Y}$ ($\mathcal{X}$: feature, $\mathcal{Y}$: response) to predict Y from X,
- Loss function: $\ell: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$,
- Risk: choose h to minimise the risk

$$R(h) = \int_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \ell(h(x), y)\, dP_0(x,y) = \mathbb{E}[\ell(h(X), Y)]$$

	classification	Regression
$\mathcal{Y}$	categorical	$\mathbb{R}$
$\ell(h(x), y)$	$\mathbb{1}(h(x) \neq y)$	$(h(x)-y)^2$
$R(h)$	$\mathbb{P}(h(X) \neq Y)$	$h(x) = \mathbb{E}[Y \mid X = x]$

$\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$

② **The Bayes classifier:** $\mathcal{Y} = \{0,1\}$, characterise $h = \underset{h}{\operatorname{argmin}}\, \mathbb{P}(h(X) \neq Y)$ (minimum a-posteriori estimator)

- Prop 1: The Bayes classifier $h_*(x) = \underset{h}{\operatorname{argmin}}\, \mathbb{E}[\ell(h(X), Y)] = \begin{cases} 1 & \text{if } \eta(x) > \frac{1}{2} \\ 0 & \text{if } \eta(x) < \frac{1}{2} \end{cases}$

Pf: $R(h) = \mathbb{E}[\mathbb{1}\{h(X) \neq Y\}] = \mathbb{E}[\underbrace{\mathbb{E}[\mathbb{1}\{h(X) \neq Y\} \mid X]}_{(*)}]$ [Tower property]

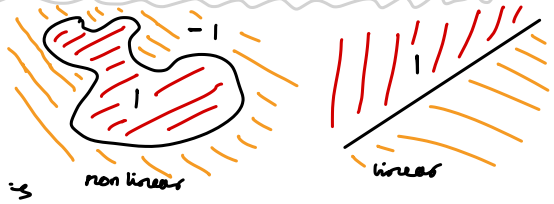
$(*) = \mathbb{E}[\mathbb{1}\{h(X)=0, Y=1\} + \mathbb{1}\{h(X)=1, Y=0\} \mid X]$ [cases on $\neq$]

$= \mathbb{1}\{h(X)=0\}\underbrace{\mathbb{E}[Y=1 \mid X]}_{\eta(X)} + \mathbb{1}\{h(X)=1\}\underbrace{\mathbb{P}(Y=0 \mid X)}_{1-\eta(X)}$ [taking out known]

If $\eta(x) > 1-\eta(x) \Rightarrow h(x) = 0$ minimises
If $\eta(x) < 1-\eta(x) \Rightarrow h(x) = 1$ minimises
} cases $\Rightarrow h = h_*$ ■

Motivation: in practice, we observe data $\{(X_i, Y_i)\}_{i=1}^n$ w/ P_0 probability distribution unknown! Instead of estimating P_0 from data, we'll define $\mathcal{H}$ & pick best $h \in \mathcal{H}$.

e.g. $\mathcal{H} = \{h : h(x) = \operatorname{sgn}(b + \sum_{j=1}^n w_j \varphi_j(x)),\ w \in \mathbb{R}^d,\ b \in \mathbb{R}\}$ (given set of basis functions $\varphi_j : \mathcal{X} \to \mathbb{R}$)



③ **Empirical Risk Minimisation & Hypothesis classes**

- Suppose $(X_i, Y_i) \overset{iid}{\sim} P_0$ for $i \in \{1, \dots, n\}$. The empirical risk is

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^n \ell(h(X_i), Y_i), \qquad \hat{h} = \underbrace{\underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \hat{R}(h)}_{\text{empirical risk minimiser}}, \qquad \text{clearly } \mathbb{E}[\hat{R}(h)] = R(h)$$

Pf: $\mathbb{E}[\hat{R}(h)] = \mathbb{E}[\frac{1}{n}\sum_{i=1}^n \ell(h(X_i), Y_i)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\ell(h(X_i), Y_i)] = \frac{1}{n} \cdot n \cdot R(h) = R(h)$

- E.g. linear regression
  - $\mathcal{X} = \mathbb{R}^p$, $\mathcal{Y} = \mathbb{R}$, $\ell(z,y) = (z-y)^2$, $\mathcal{H} = \{x \mapsto b + x^T w,\ w \in \mathbb{R}^p, b \in \mathbb{R}\}$
  - $\hat{h}(x) = x^T \hat{w} + \hat{b}$ w/ $(\hat{b}, \hat{w}) = \underset{w,b}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (Y_i - X_i^T w - b)^2$
- E.g. k-nearest neighbours
  - $\mathcal{Y} = \{-1, 1\}$, consider k nearest training pts to $\bar{x}$ and assign the most common label.
  - $d_{\bar{x}} = (\|\bar{x} - X_i\|)_{i=1}^n$ (n-vector of distances), $I_{\bar{x}}$ indexes k smallest entries in $d_{\bar{x}}$, $h(\bar{x}) = \operatorname{sgn}\left(\frac{1}{k}\sum_{i \in I_{\bar{x}}} Y_i\right)$ (vote on class by k-nearest nbs)

larger hypothesis class ⇒ better bayes classifier approximation

④ **Bias-variance Trade off:** Larger $\mathcal{H}$ ⇒ best h_* estimator, but more sensitivity to initial data.

- Regression Example
  - Dataset $D = \{(X_i, Y_i)\}_{i=1}^n$ drawn iid from P_0,
  - $\bar{y}(x) = \mathbb{E}[Y \mid X = x]$ expected output
  - $h_D(x)$ is the hypothesis learnt from D
  - $\bar{h}(x) := \mathbb{E}_{D \sim P_0^n}[h_D(x)]$ expected classifier

Note: cross term disappears: $\mathbb{E}[\mathbb{E}[(h_D(x) - \bar{y}(x))(\bar{y}(x) - Y) \mid x]] = 0$ by tower property & taking out known

expected test error: $\mathbb{E}[(h_D(x) - Y)^2] = \mathbb{E}[(h_D(x) - \bar{y}(x) + \bar{y}(x) - Y)^2] = \underbrace{\mathbb{E}[(h_D(x) - \bar{y}(x))^2]}_{(*)} + \underbrace{\mathbb{E}[(\bar{y}(x) - Y)^2]}_{\text{noise}}$

$(*) = \mathbb{E}[(h_D(x) - \bar{h}(x) + \bar{h}(x) - \bar{y}(x))^2] = \underbrace{\mathbb{E}[(h_D(x) - \bar{h}(x))^2]}_{\text{variance}} + \underbrace{\mathbb{E}[(\bar{h}(x) - \bar{y}(x))^2]}_{\text{bias}} + \underbrace{2\mathbb{E}[(h_D(x) - \bar{h}(x))(\bar{h}(x) - \bar{y}(x))]}_{=0}$

($\mathbb{E}_D(h_D(x)) = \bar{h}(x)$)

$\mathcal{H}$ ↑ complex ⇒ ↑ variance, $\mathcal{H}$ ↓ complex ⇒ ↑ bias

- Notes do same w/ k-NN [skipped as too hard & exam]

- **Cross Validation:** split dataset $D = \{(x_i, y_i)\}_{i=1}^n$ → D_{test}, D_{train}

Pick the estimator which minimizes the empirical risk.
↳ V-fold splits data into V sets, pick $h_\theta = \arg\min_{h_\theta} \frac{1}{V}\sum_{i=1}^{V} E_{i,h_\theta}$ . . .

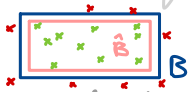
⑥ **Excess risk:** Given an estimator $\hat{h}$, the excess risk is given by

$$\mathcal{E}(\hat{h}) = R(\hat{h}) - R(h_*)$$

(h_*: Bayes classifier "best you can do")

Idea: will theoretically analyse this for different classes:
- How does $\mathcal{H}$ complexity affect $\mathcal{E}$?
- How does # data pts affect $\mathcal{E}$?

Note: "No free lunch" thms ⇒ impossible to learn all probability distributions w/out some assumptions.



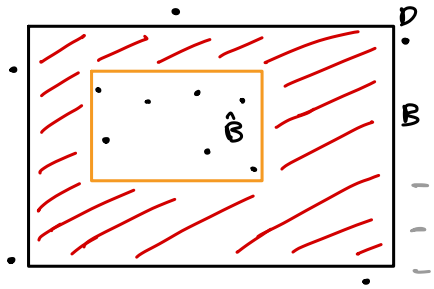
- **Example (Learning Rectangles):** $X = \mathbb{R}^2$, $Y = \{0,1\}$, goal: learn $h(x) = \mathbb{1}_B(x) = \begin{cases} 1 & \text{if } x \in B \\ 0 & \text{if } x \notin B \end{cases}$

Given iid data $(X_i, Y_i) \in X \times Y$ w/ $Y_i = \begin{cases} 1 & \text{if } X_i \in B \\ 0 & \text{if } X_i \notin B \end{cases}$, $\hat{h} = \mathbb{1}_{\hat{B}}$ where $\hat{B}$ is the smallest rectangle containing pts labelled as inside B. Note: $R(h_*) = 0$ ∵ $h_* = \mathbb{1}_B$ is the Bayes classifier

$B = [a,b] \times [c,d]$

$$\mathcal{E}(\hat{h}) = R(\hat{h}) = \mathbb{P}(\hat{h}(x) \neq \mathbb{1}_B(x)) = \mathbb{P}(X \in B \setminus \hat{B}) + \underbrace{\mathbb{P}(X \in \hat{B} \setminus B)}_{=0 \because \hat{B} \subset B}$$

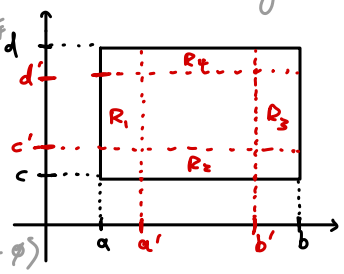
PAC: approximately correct, probably



Q: How many samples to ensure $\mathbb{P}(R(\hat{h}) \leq \varepsilon) \geq 1-\delta$?

Assume $\mathbb{P}(X \in B) > \varepsilon$ o/w, $R(\hat{h}) = \mathbb{P}(X \in B \setminus \hat{B}) \leq \varepsilon$ trivially!

- Define $R_1 = [a, a'] \times [c, d]$, a' smallest s.t. $\mathbb{P}(X \in R_1) \geq \varepsilon/4$
- Def $R_2 = [a, b] \times [c, c']$, c' " s.t. $\mathbb{P}(X \in R_2) \geq \varepsilon/4$
- Def $R_3 = [b', b] \times [c, d]$, b' largest s.t. $\mathbb{P}(X \in R_3) \geq \varepsilon/4$
- Def $R_4 = [a, b] \times [d', d]$, d' largest s.t. $\mathbb{P}(X \in R_4) \geq \varepsilon/4$



Note that

$$\text{if } (\hat{B} \cap R_i \neq \emptyset \ \forall i) \Rightarrow \mathbb{P}(X \in B \setminus \hat{B}) \leq \varepsilon \text{ so } \mathbb{P}(R(\hat{h}) \leq \varepsilon) \geq \mathbb{P}(\hat{B} \cap R_i \neq \emptyset \ \forall i)$$

$$\{\hat{B} \cap R_i \neq \emptyset \ \forall i\} \supset \{R(\hat{h}) \leq \varepsilon\}$$

$$= 1 - \mathbb{P}(\exists i \text{ s.t. } \hat{B} \cap R_i = \emptyset) \geq 1 - \sum_{i=1}^{4} \mathbb{P}(\hat{B} \cap R_i = \emptyset) \quad \text{[union bd]}$$

$$\mathbb{P}(\hat{B} \cap R_i = \emptyset) = \prod_{j=1}^{n} \mathbb{P}(X_j \notin R_i) \leq (1 - \varepsilon/4)^n \leq e^{-\varepsilon n/4} \text{ so can put together}$$

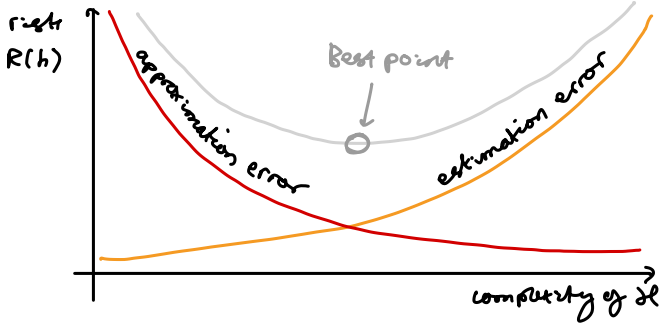
$$\mathbb{P}(R(\hat{h}) \leq \varepsilon) \geq 1 - 4e^{-\varepsilon n/4} \geq 1 - \delta \iff \delta > 4e^{-\varepsilon n/4} \iff -\log\left(\tfrac{4}{\delta}\right) \geq -\tfrac{\varepsilon n}{4} \iff n \geq \tfrac{4}{\varepsilon}\log\left(\tfrac{4}{\delta}\right)$$

Note: As $h_* = \mathbb{1}_B$, $R(h_*) = 0$, now consider $h_* \notin \mathcal{H}$ so $R(h_*) > 0$

We want to minimize the excess risk

- **Decomposition of Excess Risk:** excess risk

$$R(\hat{h}) - R(h_*) = \underbrace{R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h)}_{\text{Estimation Error [variance]}} + \underbrace{\inf_{h \in \mathcal{H}} R(h) - R(h_*)}_{\text{Approximation error [Bias]}}$$



- **Approximation error:** Need assumptions on h_* & P_0

e.g. $z \mapsto \ell(y, z)$ L-lipschitz cts. Then

$$R(h) - R(h') = \mathbb{E}[\ell(h(x), Y) - \ell(h'(x), Y)] \leq \mathbb{E}[|h'(x) - h(x)|]$$

Assume $h_*(x) = w_*^T \varphi(x)$ for $\varphi(x) = (1, x, \ldots, x_p, x_1^2, x_1 x_2, \ldots)$

and $\mathcal{H} = \{h_w(x) = w^T \varphi(x) : \|w\| \leq k\}$. Then

$$\inf_{h \in \mathcal{H}} R(h) - R(h_*) \leq L \inf_{\|w\| \leq k} \mathbb{E}[|\langle \varphi(x), w - w_* \rangle|] \quad \text{(zero if } \|w_*\| \leq k\text{)}$$

$$\leq L\, \mathbb{E}\|\varphi(x)\| \max\{\|w_*\| - k, 0\}$$

$$:= \inf_{\|w\| \leq k} \{\|w - w^*\|\}$$

⑦ **The estimation Error:** consider for ERM $\hat{h}$.

letting $\bar{h} = \arg\min_{h \in \mathcal{H}} R(h)$, note that $\hat{R}(\hat{h})$ is as small as possible: $\hat{h}$ is ERM, ≤ 0

$$R(\hat{h}) - R(\bar{h}) = R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\hat{h}) - \hat{R}(\bar{h}) + \hat{R}(\bar{h}) - R(\bar{h}) \quad (*)$$

estimation error for $\hat{h}$

$$\leq \sup_{h \in \mathcal{H}} R(h) - \hat{R}(h) + \hat{R}(\bar{h}) - R(\bar{h}) \quad \text{[relaxing conditions]}$$

Note: CLT is asymptotic result so no quantative bounds... we need concentration inequalities!

$$\leq 2 \sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)|$$

GOAL: $\mathbb{P}\left(\sup_{h \in \mathcal{H}} |R(h) - \hat{R}(h)| > \varepsilon\right) \leq \delta$

Precisely control the estimation error of other hypotheses does $\mathcal{H}$

- **Tools from Probability**
  - (a) Markov: $\mathbb{P}(w \geq t) \leq \frac{\mathbb{E}[w]}{t}$
  - (b) Chernoff: $\mathbb{P}(w \geq t) \leq \inf_{\alpha > 0} e^{-\alpha t}\, \mathbb{E}[\exp(\alpha w)]$
  - (c) Sub-Gaussian: w sub-gaussian w/ param σ if $\mathbb{E}\left[e^{\alpha(w - \mathbb{E}[w])}\right] \leq e^{\alpha^2\sigma^2/2}$ $\forall \alpha \in \mathbb{R}$
    - ↳ If w sub-gaussian, $\mathbb{P}(w - \mathbb{E}[w] \geq t) \leq e^{-t^2/2\sigma^2}$ $\forall t \geq 0$
    - ↳ If $w \in [a,b] \Rightarrow w$ sub-gaussian w/ $\sigma = \frac{b-a}{2}$
    - ↳ $w_1, \dots, w_n$ independent. w_i subgaussian w/ param σ_i. Then $\forall \gamma \in \mathbb{R}^n$, $\sum_{i=1}^n \gamma_i w_i$ sub-g. w/ param $\left(\sum \gamma_i^2 \sigma_i^2\right)^{\frac{1}{2}}$
  - (d) Hoeffding: $w_1, \dots, w_n$ independent & $a_i \leq w_i \leq b_i$ $\forall i$. Then $\forall t \geq 0$ $\bar{z} = \frac{1}{n}\sum_{i=1}^n w_i$

$$\mathbb{P}(|\bar{z} - \mathbb{E}[\bar{z}]| \geq t) \leq 2\exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

- **Finite Hypothesis classes:** (Thm 7) Assume $\ell(h(x), y) \in [0, L]$, $\forall h \in \mathcal{H}$. $\bar{h} = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$, $|\mathcal{H}| < \infty$. Then w/ prob. at least $1-\delta$, the ERM $\hat{h}$ satisfies

(approximation error for $\hat{h}$)

$$R(\hat{h}) - R(\bar{h}) \leq L\sqrt{\frac{2\log(|\mathcal{H}|) + \log(\delta^{-1})}{n}}$$

Pf: If $\hat{h} = \bar{h} \Rightarrow R(\hat{h}) - R(\bar{h}) = 0 \Rightarrow \mathbb{P}(R(\hat{h}) - R(\bar{h}) > t) = \mathbb{P}(R(\hat{h}) - R(\bar{h}) > t,\ \hat{h} \neq \bar{h})$

By (*) $\mathbb{P}(R(\hat{h}) - R(\bar{h}) > t) \leq \mathbb{P}(R(\hat{h}) - \hat{R}(\hat{h}) + \hat{R}(\bar{h}) - R(\bar{h}) \geq t,\ \hat{h} \neq \bar{h})$

$$\leq \underbrace{\mathbb{P}(R(\hat{h}) - \hat{R}(\hat{h}) > t/2,\ \hat{h} \neq \bar{h})}_{(**)} + \underbrace{\mathbb{P}(\hat{R}(\bar{h}) - R(\bar{h}) > t/2)}_{(***)}$$

Bound (***) w/ Hoeffding. $\hat{R}(\bar{h}) - R(\bar{h}) = \frac{1}{n}\sum_{i=1}^n \ell(\bar{h}(x_i), y_i) - \mathbb{E}[\ell(\bar{h}(x_i), y_i)]$

So

$$\mathbb{P}(\hat{R}(\bar{h}) - R(\bar{h}) > t/2) \leq \exp\left(-\frac{2n^2(t/2)^2}{nL^2}\right) = e^{-\frac{nt^2}{2L^2}}$$

When $\hat{h} \neq \bar{h}$, $R(\hat{h}) - \hat{R}(\hat{h}) \leq \max_{h \in \mathcal{H} \setminus \{\bar{h}\}} R(h) - \hat{R}(h)$, so

$$(**) \leq \mathbb{P}(\exists h \in \mathcal{H}\setminus\{\bar{h}\} : R(h) - \hat{R}(h) > t/2) \leq \sum_{h \in \mathcal{H}\setminus\{\bar{h}\}} \mathbb{P}(R(h) - \hat{R}(h) > t/2) \quad \text{[union bd]}$$

$$\leq (|\mathcal{H}| - 1)\exp\left(-\frac{nt^2}{2L^2}\right) \quad \text{[Hoeffding]}$$

$$\therefore \mathbb{P}(R(\hat{h}) - R(\bar{h}) > t) \leq |\mathcal{H}|\exp\left(-\frac{nt^2}{2L^2}\right) \leq \delta$$

$$\Leftrightarrow -\log\left(\frac{|\mathcal{H}|}{\delta}\right) \geq -\frac{nt^2}{2L^2}$$

$$\Leftrightarrow t^2 \geq 2L^2\log(|\mathcal{H}|/\delta)/n$$

Note: $\mathbb{P}(\forall h \in \mathcal{H}, R(h) - \hat{R}(h) \leq t) = 1 - \mathbb{P}(\exists h \in \mathcal{H}, R(h) - R(\hat{h}) > t)$ $\quad \delta = |\mathcal{H}|e^{-\frac{2t^2n}{L^2}}$

can use this idea in model selection (how to chose $\mathcal{H}$) w/ techniques known as structural risk minimization

$$\geq 1 - \sum_{h \in \mathcal{H}} \mathbb{P}(R(h) - R(\hat{h}) > t) \qquad \log\left(\frac{\delta}{|\mathcal{H}|}\right) = -\frac{2t^2n}{L} \Rightarrow t^2 = \frac{L\log(|\mathcal{H}|/\delta)}{2n}$$

$$\geq 1 - \underbrace{|\mathcal{H}|\exp\left(-\frac{t^2n^2}{nL^2}\right)}_{\delta} \quad \text{[Hoeffding]}$$

$\mathbb{E}[\ell(h(x), Y)]$

$$\therefore \forall h \in \mathcal{H},\ \mathbb{P}\left(R(h) < \hat{R}(h) + \sqrt{\frac{L^2\log(|\mathcal{H}|/\delta)}{2n}}\right) \geq 1 - \delta \quad \text{[holds for } h = \hat{h}\text{]}$$

($R(h)$: cannot be computed; $\hat{R}(h)$: evaluate w/ data)

Example: $\mathcal{X} = [0,1)^2$ partitioned into m^2 disjoint squares length $\frac{1}{m}$ each. $|\mathcal{H}| = 2^{m^2}$ & apply

---

- **Infinite Hypothesis classes:** $|\mathcal{H}| = \infty$, As $R(\hat{h}) - R(\bar{h}) \leq R(\hat{h}) - \hat{R}(\hat{h}) - \hat{R}(\bar{h}) - R(\bar{h})$

$\mathbb{E}[R(\hat{h}) - R(\bar{h})] \leq \mathbb{E}[G]$ $\quad G := \sup_{h \in \mathcal{H}} R(h) - \hat{R}(h)$

(take expectation)

GOAL: bound $\mathbb{E}[G]$ & hence bound the expected risk

$Y = \{-1, 1\} \Rightarrow$ BINARY CLASSIFICATION

can write this for ease...

Set $z_i = (x_i, y_i)$, consider $\mathcal{F} = \{(x,y) \mapsto \ell(h(x), y) : h \in \mathcal{H}\}$, $G = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n \mathbb{E}[f(z_i)] - f(z_i)$

Rademacher complexity: $\mathcal{F}$ class of $f: \mathcal{Z} \to \mathbb{R}$, $z_1, \dots, z_n \in \mathcal{Z}$

(a) $\mathcal{F}(z_{1:n}) = \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F}\} \subset \mathbb{R}^n$ is a collection of vectors giving the behaviours of $\mathcal{F}$ on $z_{1:n}$.

(b) The empirical rademacher complexity is

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) := \frac{1}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i f(z_i)\right]$$

"how closely aligned is your classifier on your dataset?"

w/ $(\varepsilon_i)_{i=1}^n$ iid rademacher. This quantifies how close $f(z_{1:n})$ is to random labels.

(c) Take $z_1, \dots, z_n$ as R.V. The rademacher complexity is

$$\mathcal{R}(\mathcal{F}) := \mathbb{E}[\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))]$$

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) = \frac{1}{n} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_i \varepsilon_i f(z_i) \mid z_{1:n}\right]$$

Some examples... pick sup f carefully based on situation.

Idea: bd $\hat{\mathcal{R}}(\mathcal{F}(z_{1:n}))$ uniformly in $z_{1:n}$

↳ **Thm 8:** In setting above, $f: \mathcal{Z} \to \mathbb{R}$

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n (\mathbb{E}[f(z_i)] - f(z_i))\right] \le 2\mathcal{R}_n(\mathcal{F})$$

(LHS: depends on z_i distributions... complex; RHS: independent of P_0)

Proofs are non-examinable.

Pf: immediate consequence

↳ **Thm 9:** $\mathcal{F} := \{(x,y) \mapsto \ell(h(x), y) ; h \in \mathcal{H}\}$. Then

$$\mathbb{E}[R(\hat{h}) - R(\bar{h})] \le 2\mathcal{R}_n(\mathcal{F})$$

MAIN RESULT OF SECTION

THE ESTIMATION ERROR IS CONTROLLED BY THE RADEMACHER COMPLEXITY

Example 9 (linear models): If $\ell(\cdot, y)$ is G-lipschitz, then $\mathcal{F} = \{\ell(h(x), y), h \in \mathcal{H}\}$ satisfies $\mathcal{R}_n(\mathcal{F}) \le G\mathcal{R}_n(\mathcal{H})$. Take $\mathcal{H} = \{h_w(x) = w^T \varphi(x) ; \|w\| \le k\}$. Then ...

End up w/

$$\mathbb{E}[R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h)] \le \frac{k \sqrt{\mathbb{E}[\|\varphi(x)\|^2]}}{\sqrt{n}}$$

long & hard... won't come up...

independent of φ dimension!

VC Dimension: Bounding expected risk → bounding rademacher complexity

(study w/ VC dimension)

Idea: given data $z_{1:n}$, count behaviours of $\mathcal{F}$: $|\mathcal{F}(z_{1:n})|$

Note: $|\mathcal{F}(z_{1:n})| \le |\mathcal{H}(z_{1:n})|$ ∵ have an injection $\mathcal{F} \to \mathcal{H}$

$(\ell(h(x_i), y_i))_{i=1}^n \ne (\ell(h'(x_i), y_i))_{i=1}^n$

$\Downarrow$

$h(x_i) \ne h'(x_i)$

Lemma 11: $w_1, \dots, w_d$ (possibly σ-) mean zero s.g. $\Rightarrow \mathbb{E}[\max_j w_j] \le \sigma\sqrt{2\log(d)}$

Pf: use $m \in \mathcal{F}$, add more terms & optimize

THIS IS WHAT VC DIM ACHIEVES

Lemma 10 (Massart's lemma): $\mathcal{F} = \{(x,y) \mapsto \ell(h(x), y), h \in \mathcal{H}\}$ (ℓ loss compared w/ h), ℓ take values in $[0,1]$. Then

$$\hat{\mathcal{R}}(\mathcal{F}(z_{1:n})) \le \sqrt{\frac{2\log|\mathcal{F}(z_{1:n})|}{n}} \le \sqrt{\frac{2\log|\mathcal{H}(x_{1:n})|}{n}}$$

order $\frac{1}{\sqrt{n}}$ convergence just like the finite case for $|\mathcal{H}|$

Pf: By def have $\mathbb{E}[\sup_{v \in \mathcal{F}(z_{1:n})} \frac{1}{n} \sum_{i=1}^n \varepsilon_i v_i]$ (sum of subgaussian R.V.), so need exactly $\mathbb{E}[\max_{j=1,\dots,d} w_j] \le \sigma\sqrt{2\log(d)}$

Let $d = |\mathcal{F}(z_{1:n})|$ (finite family)

Let $\mathcal{F}' = \{f_1, \dots, f_d\}$ s.t. $\mathcal{F}'(z_{1:n}) = \mathcal{F}(z_{1:n})$.

So let $w_j = \frac{1}{n} \sum_{i=1}^n f(z_i)\varepsilon_i \Rightarrow \mathbb{E}[\max_{j=1,\dots,d} w_j] \le \sqrt{\frac{2\log(d)}{n}}$ (n ← parameter)

Rademacher R.V. sub-gaussian w/ $\sigma = 1$ so apply $\frac{1}{n}\sum_{i=1}^n f(z_i)\varepsilon_i$ s.g. w/ $\sigma = \sqrt{\sum_{i=1}^n \frac{f(z_i)^2}{n^2}} \le \frac{1}{\sqrt{n}}$

Note: $\mathbb{E}[R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h)] \le 2\mathcal{R}_n(\mathcal{F}) \le 2\mathbb{E}\sqrt{\frac{2\log|\mathcal{H}(x_{1:n})|}{n}}$ (estimation error)

$\Rightarrow$ top grows slower than bottom to be useful $\Rightarrow |\mathcal{H}(x_{1:n})|$ slower than exponential e.g. polynomial

If $Y \in \{-1, 1\}$, $\mathcal{H}(x_{1:n}) = \{(h(x_i))_{i=1}^n : h \in \mathcal{H}\} \Rightarrow |\mathcal{H}(x_{1:n})| \le 2^n$ BAD! (each entry ± 1)

GOAL: characterise classes of functions where $|\mathcal{H}(x_{1:n})| << 2^n \Rightarrow$ Massart gives ✓ bound.

Otherwise, estimation error unbdd ⇒ useless class of functions!

Def: Let $\mathcal{H}$ be a class of functions $h: \mathcal{X} \to \{a, b\}$, $a \ne b$. $|\mathcal{H}| \ge 2$.

VC Dimension definition

- $\mathcal{H}$ shatters $x_{1:n} \in \mathcal{X}^n$ if $|\mathcal{H}(x_{1:n})| = 2^n$ "every possible labelling is achieved on $x_{1:n}$"
- $\Pi_{\mathcal{H}}(n) = \max_{x_{1:n} \in \mathcal{X}^n} |\mathcal{H}(x_{1:n})|$ is the shattering coefficient / growth function "max # behaviours on n data points using $\mathcal{H}$"
- The vc dimension of $\mathcal{H}$, $VC(\mathcal{H}) = \sup\{n \in \mathbb{N} : \Pi_{\mathcal{H}}(n) = 2^n\}$ "largest # data points s.t. $x_{1:n}$ is shattered"

largest # datapoints that achieve every possible labelling

To check $VC(\mathcal{H}) = n$: ① Find *some* $x_{1:n}$ that can be shattered by $\mathcal{H}$ (any combo works) ② Show *no set* of $n+1$ pts can be shattered (so show how any arbitrary $n+1$ pts can be shattered by $\mathcal{H}$)

Example: $X = \mathbb{R}$, $Y = \{0,1\}$, $\mathcal{H} = \{\mathbb{1}_{[a,b)} : a, b \in \mathbb{R}\}$. Given data $x_1 < x_2 < \dots < x_n$, how many distinct vectors in $\mathcal{H}(x_{1:n})$?

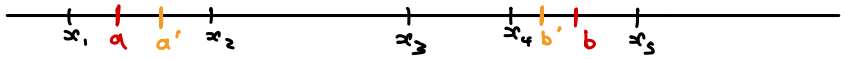
If a & a' in the same interval $(x_i, x_{i+1}]$ $\Rightarrow \left(\mathbb{1}_{[a,b)}^{(x_j)}\right)_j = \left(\mathbb{1}_{[a',b')}^{(x_j)}\right)_j$

If b & b' in the same interval $(x_j, x_{j+1}]$

There are $n+1$ intervals $(-\infty, x_1], (x_1, x_2], \dots, (x_{n-1}, x_n], (x_n, \infty)$

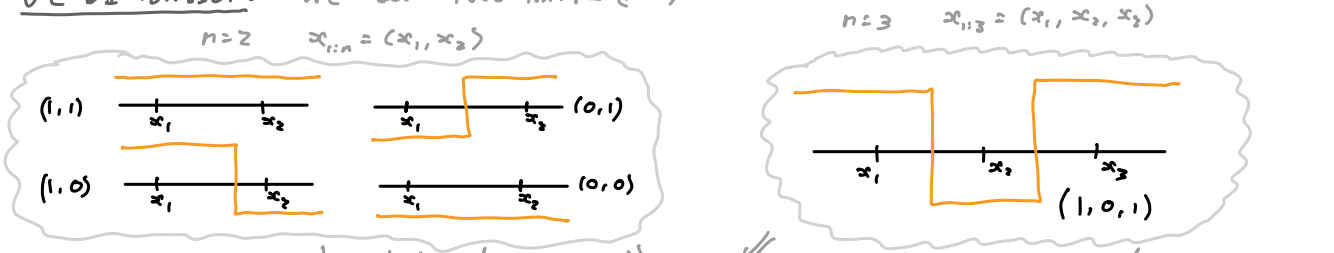
so we get the same behaviour.

Note: To show VC dim



So as an upper bd, $n+1$ intervals to place a, $n+1$ intervals to place b $\Rightarrow |\mathcal{H}(x_{1:n})| \leq (n+1)^2$

VC DIMENSION: we get $|\mathcal{H}(x_{1:n})| \leq (n+1)^{VC(\mathcal{H})}$



$x_{1:2}$ can be shattered ↳ all possible labellings achieved

$\Rightarrow VC(\mathcal{H}) = 2$

$x_{1:3}$ cannot be shattered ↳ (1,0,1) labelling not achieved in $\mathcal{H}$

Pf: non-examinable.

Sauer-Shelah lemma: $\mathcal{H}$ a class w/ finite vc dimension d, then $\Pi_{\mathcal{H}}(n) \leq (n+1)^d$

So, overall, $|\mathcal{H}(z_{1:n})| \leq \Pi_{\mathcal{H}}(n) \leq (n+1)^d$ where $d = VC(\mathcal{H})$, so

$$\mathbb{E}[R(\hat{h}) - \inf_{h \in \mathcal{H}} R(h)] \leq 2\mathcal{R}(\mathcal{F}) \leq 2\sqrt{\frac{2\,VC(\mathcal{H})\log(n+1)}{n}} \xrightarrow{n\to\infty} 0$$

(Thm 9; Massart's lemma)

expected estimation error

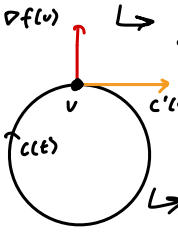
CHAPTER TWO

Optimisation

$\operatorname{argmin}(f) = \{w \in \mathbb{R}^p : f(w) = \inf_{w \in \mathbb{R}^p} f(w)\}$

⑧ Unconstrained Optimisation: $f: \mathbb{R}^p \to \mathbb{R}$, goal: $\inf_{w \in \mathbb{R}^p} f(w)$

- $w \in \mathbb{R}^p$ is a
  - ↳ global minimiser if $\forall w \in \mathbb{R}^p \quad f(w^*) \leq f(w)$
  - ↳ local " " $\exists$ open neighbourhood U of w^* s.t. $f(w^*) \leq f(w) \quad \forall w \in U$
  - ↳ strict local " if " " " " " " $f(w^*) < f(w) \quad \forall w \in U \setminus \{w^*\}$
  - ↳ isolated local minimiser if " " " " " in which w^* is the only local minimiser
- A set C is convex if $\forall w, v \in C$, $\forall \lambda \in [0,1]$, $\lambda v + (1-\lambda)w \in C$
- A function $f: S \to \mathbb{R}$ convex if $f(\lambda v + (1-\lambda)w) \overset{\text{strict!}}{\leq} \lambda f(v) + (1-\lambda) f(w) \quad \forall v, w \in S, \forall \lambda \in [0,1]$
  - ↳ concave if $-f$ convex
- $f: \mathbb{R}^p \to \mathbb{R}$ convex. Then any local minimiser is a global minimiser. If strictly convex ↑ unique
  - Pf: probs won't come up...
- Recall that $f: \mathbb{R}^p \to \mathbb{R}$ diff $\Rightarrow f(w+v) = f(w) + \langle \nabla f(w), v \rangle + o(\|v\|)$ [Taylor]
- Characterisation of convexity via differentiability:
  - (i) If f diff. f convex $\iff \forall w, v \in \mathbb{R}^p \quad f(w) \geq f(v) + \nabla f(v)^T (w - v)$
  - (ii) If f twice diff f convex $\iff$ Hessian $\nabla^2 f(v)$ positive semi-definite $\forall v \in \mathbb{R}^p$ ($v^T A v \geq 0$ $\forall v$)
  - Pf: probs won't come up...
- First order optimality condition: f convex & diff. Then $w^* \in \operatorname{argmin}(f) \iff \nabla f(w^*) = 0$
  - Pf: '$\Rightarrow$' well known, '$\Leftarrow$' convexity $\Rightarrow f(w) \geq f(w^*) + \overbrace{\nabla f(w^*)}^{=0}(w - w^*) = f(w^*) \Rightarrow w^*$ minimiser ■
  - Note: $f(x) = -x^2$, $\nabla f(0) = 0$ but 0 maximiser
    $f(x) = x^3$ $\nabla f(0) = 0$ but 0 not local minimiser
- Example (Least Squares): $f(w) = \|Aw - b\|^2$ $A \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^n$. As $f(x+\varepsilon) = f(x) + \nabla f(x)^T \varepsilon$
  $f(w+\varepsilon) = \|Aw - b + A\varepsilon\|^2 = \|Aw-b\|^2 + 2\langle Aw-b, A\varepsilon\rangle + \|A\varepsilon\|^2 \Rightarrow \nabla f(w) = 2A^T(Aw - b)$
  Convex $\Rightarrow w^*$ minimiser iff $A^TAw^* = A^Tb$. If A^TA invertible ($\ker(A^TA) = \{0\}$), then $w^* = (A^TA)^{-1}A^Tb$. [note $\nabla^2 f = 2A^TA$ is useful for time dep]
- Gradient descent: can't compute $w^* \in \operatorname{argmin}(f)$ in closed form $\Rightarrow$ construct $w_t \xrightarrow{t\to\infty} w^*$ [inverting big matrices]



↳ **Direction of steepest descent**: Require $\langle \nabla f(x_0), p \rangle < 0$ as $-\nabla f(x_0)$ is the direction of steepest descent. (direction) NOTE: $\nabla f(v)$ is orthogonal to level sets.

GRADIENT DESCENT

Initialise w/ $w_0 \in \mathbb{R}^p$. For $k = 0, 1, 2, 3, \ldots$ $w_{k+1} = w_k - \tau_k \nabla f(w_k)$ (learning rate / step-size)

↳ **Choosing the stepsize**: Not too small [converge ASAP], not too big [increase f, not converge]

Ⓐ Greedy choice: $\tau = \operatorname{argmin}_{\tau} h(\tau)$, $h(\tau) = f(w_k - \tau \nabla f(w_k))$

compute $h'(\tau)$ set $h'(\tau_k) = 0$

$h(\tau + \varepsilon) = f(w_k - \tau \nabla f(w_k) - \varepsilon \nabla f(w_k))$

$= f(w_k - \tau \nabla f(w_k)) + \varepsilon \nabla f(w_k)^T \nabla f(w_k - \tau_k \nabla f(w_k)) + o(\varepsilon)$

$\Rightarrow h'(\tau) = \nabla f(w_k)^T \nabla f(w_k - \tau_k \nabla f(w_k)) = 0 \Rightarrow \nabla f(w_k)^T \nabla f(w_{k+1}) = 0$

As $\nabla f(w_k) \nabla f(w_{k+1}) = 0 \Rightarrow$ orthogonal gradient descent directions $\Rightarrow$ zig-zag

Ⓑ Armijo rule: Find τ that sufficiently decreases f. Some backtracting line search.

Example (least squares): Greedy choice $\Rightarrow 0 = \nabla f(w_k)^T \nabla f(w_{k+1}) = \langle r_k, r_{k+1} \rangle$

$w_{k+1} = w_k - \tau_k r_k$, $r_k = \nabla f(w_k) = A^T(Aw_k - b) \Rightarrow \ldots \Rightarrow \tau_k = \frac{\|r_k\|^2}{\|Ar_k\|^2}$

↳ **Convergence Analysis for GD**: [don't need proofs for exam, state results here]...

- If $0 < \tau_{min} \le \tau_k \le \tau_{max} < \frac{2}{L}$. Then $\exists \rho \in [0,1)$ s.t. $\|w_k - w^*\| \le \rho^k \|w_0 - w^*\|$ [convex problem]
  ↳ ie w_k converges linearly to w^* [L largest eigenvalue for C: $f(w) := \frac{1}{2}\langle Cw, w \rangle - \langle w, b \rangle$]
- $f : S \longrightarrow \mathbb{R}$ diff, $S \subset \mathbb{R}^p$ convex
  (i) f is μ-strongly convex if $\forall x, x' \in S$, $\langle \nabla f(x) - \nabla f(x'), x - x' \rangle \ge \mu \|x - x'\|^2$
  (ii) f is L-lipschitz smooth if " $\|\nabla f(x) - \nabla f(x')\| \le L \|x - x'\|$
  ↳ These imply f can be lower & upper bded by quadratics...

Main result on convergence of GD

When f satisfies (i)&(ii) $\Rightarrow$ linear rate of convergence $\|w_k - w^*\| \le \rho^k \|w_0 - w^*\|$

When f only (ii) (lipschitz smooth) $\Rightarrow$ $O(\frac{1}{k})$ convergence [slower] $f(w_k) - f(w^*) \le \frac{C}{k}$

When f lipschitz cts $\Rightarrow$ $O(\frac{1}{\sqrt{k}})$ convergence [even worse] . . .

- **Stochastic Gradient Descent**: what if for each iteration k, draw an index i uniformly at random from $\{1, \ldots, n\}$? Problem: $\min_{w \in \mathbb{R}^p} f(w)$ where $f(w) = \frac{1}{n} \sum_{i=1}^n f_i(w)$

$$\mathbb{E}[\nabla f_{i_k}(w)] = \sum_{j=1}^n \nabla f_j(w) \cdot \mathbb{P}(i = j) = \frac{1}{n} \sum_{j=1}^n \nabla f_j(w) = \nabla f(w)$$

So, replace expensive gradient computation by $\nabla f_{i_k}(w)$

e.g. $\min_{w \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell(h_w(x_i), y_i)$, $\ell(z, y) = \frac{1}{2}(z - y)^2$, $f_i(w) = \frac{1}{2}(w^T x_i - y_i)^2$ — Highly relevant to ML

Stochastic Gradient Descent

Initialise $w_0 \in \mathbb{R}^p$, for $k = 0, 1, 2, \ldots$ $w_{k+1} = w_k - \nabla f_{i_k}(w_k)$ $i_k \overset{\text{unif}}{\sim} \{1, \ldots, n\}$ $\{w_k\}$ random vectors

Example: $f_i(w) = \frac{1}{2}(w^T x_i - y_i)^2 \Rightarrow \nabla f_i(w) = (w^T x_i - y_i) x_i$. If $w \in \mathbb{R}^p$ $\{x_i\}_{i=1}^n$ n data points, then each k of SGD costs $O(p)$, each k of GD $\Rightarrow \nabla f(w) \Rightarrow O(np) \Rightarrow$ n times cheaper!

↳ **Thm 19**: Suppose f μ-strongly convex, $\|\nabla f_i(w)\|^2 \le C^2$ $\forall i$. Let $\tau_k = \frac{1}{\mu(k+1)}$

Then $\mathbb{E}[\|w_k - w^*\|^2] \le \frac{R}{k+1}$ where $R = \max\{\|w_0 - w^*\|^2, \frac{C^2}{\mu^2}\}$

Remarks:

- SGD much slower than GD. Only weakly benefits from strong convexity & no linear rate
- Advantage: SGD makes progress towards w^* w/out full pass through n data points. (w/ large datasets, you might only make 1 full pass).

---

⑨ **Constrained Optimisation**: Consider $\min_{w \in \mathbb{R}^p} f_0(w)$ subject to $Aw = b$ and $\forall i = 1, \ldots, m$, $f_i(w) \le 0$ (P)

w/ $A \in \mathbb{R}^{n \times p}$, $b \in \mathbb{R}^n$, $f_i : \mathbb{R}^p \longrightarrow \mathbb{R}$. Assume f_i convex

- Feasibility set of (P) $\mathcal{F} := \{w \in \mathbb{R}^p : \forall i \in \{1, \ldots, m\}, f_i(w) \le 0 \text{ and } Aw = b\}$ (place where all the conditions hold)
- **First order optimality condition**: Let f be convex & diff. Consider $\min_{w \in \mathcal{F}} f(w)$ for $\mathcal{F}$ convex.

Then w^* minimises $\iff$ $\forall w \in \mathcal{F}$, $\nabla f(w^*)^T (w - w^*) \ge 0$

Pf: $\Leftarrow$ $f(w) \overset{f \text{ convex}}{\geq} f(w^*) + \langle \nabla f(w^*), w - w^* \rangle \overset{\text{assumption}}{\geq} f(w^*) \Rightarrow w^*$ minimizer

$\Rightarrow$ contradiction. Assume w^* minimizer but $\nabla f(w^*)^T (w - w^*) < 0$ for some $w \in \mathcal{F}$

Let $\lambda \in [0,1]$, define $v(\lambda) = (1-\lambda) w^* + \lambda w \in \mathcal{F}$. Define $g(\lambda) = f(v(\lambda))$

$g'(\lambda) = \nabla f(v(\lambda))^T (w - w^*)$ [chain rule] $\Rightarrow g'(0) = \langle \nabla f(w^*), w - w^* \rangle < 0$ [assumption]

So g decreasing at $\lambda = 0 \Rightarrow f(v(\lambda)) < f(v(0)) = f(w^*)$ $\forall \lambda$ sufficiently small $\lightning$ ■

↳ **Duality**: Primal (P) & dual offer different perspectives to help solve

- The Lagrange function of (P) is defined $\forall w \in \mathbb{R}^p$, $\xi \in \mathbb{R}^n$, $\nu \in \mathbb{R}^m_{\geq 0}$ by

$$L(w, \xi, \nu) = \underbrace{f_0(w)}_{\text{function}} + \underbrace{\xi^T (Aw - b)}_{\text{linear}} + \underbrace{\sum_{k=1}^m \nu_k f_k(w)}_{\text{inequalities}}$$

ξ, ν are the Lagrange multipliers

- The Lagrange dual function is $D(\xi, \nu) = \inf_{w \in \mathbb{R}^p} L(w, \xi, \nu)$

note: if $w \mapsto L(w, \xi, \nu)$ unbded from below $D(\xi, \nu) = -\infty$

↳ **Property 1**: Dual function is concave: $(\xi, \nu) \mapsto D(\xi, \nu)$ linear in $\xi, \nu \Rightarrow$ ptwise infimum of linear affine so convex.

↳ **Property 2**: Dual lower bounds $\inf_{w \in \mathcal{F}} f_0(w)$

$$D(\xi, \nu) \leq L(w, \xi, \nu) = f_0(w) + \underbrace{\xi^T (Aw - b)}_{=0} + \underbrace{\sum_{k=1}^m \nu_k f_k(w)}_{\leq 0} \leq f_0(w) \quad \forall w \in \mathcal{F}$$

Taking supremum $\Rightarrow$ $D(\xi, \nu) \leq \inf_{w \in \mathbb{R}^p} f_0(w)$

- The dual problem to (P) is

$$\max_{\xi \in \mathbb{R}^n, \nu \in \mathbb{R}^m} D(\xi, \nu) \quad \text{s.t.} \quad \nu \geq 0 \qquad (D)$$

- Let (ξ^*, ν^*) maximize (D) and w^* minimize (P)
  - ↳ weak duality: $D(\xi^*, \nu^*) \leq f_0(w^*)$
  - ↳ strong duality: $D(\xi^*, \nu^*) = f_0(w^*)$

KKT CONDITIONS
- Feasibility conditions
- inequalities $\Rightarrow$ = w/ multipliers
- $\nabla L = 0$

When does strong duality hold?

Slater's constraint qualification: Assume f_i convex $\forall i \in \{0, \dots, m\}$. $\text{dom}(f_0) = \mathbb{R}^p$. If $\exists w \in \mathbb{R}^p$ s.t. $Aw = b$ and $f_i(w) < 0$ $\forall i \in \{1, \dots, m\}$. Then strong duality holds [$\exists$ pt in the INTERIOR]

If we have strong duality, how are ξ^*, ν^*, w^* related?

If w^*, ν^*, ξ^* optimize (P) and (D) & strong duality holds, then following KKT hold:

$f_i(w^*) \leq 0$, $Aw^* = b$, $\nu_i^* \geq 0$

$\nu_i^* f_i(w^*) = 0 \quad \forall i \in 1, \dots, m$

$\nabla f_0(w^*) + \sum_i \nu_i^* \nabla f_i(w^*) + A^T \xi^* = 0$

why? $f_0(w^*) = D(\xi^*, \nu^*) = \inf_{w \in \mathcal{F}} f_0(w) + \underbrace{\langle \xi^*, Aw - b \rangle}_{=0 \; \forall w \in \mathcal{F}} + \underbrace{\sum_{i=1}^m \nu_i^* f_i(w)}_{\nu_i \geq 0, f_i \leq 0} \leq f_0(w^*) + \sum_{i=1}^m \nu_i^* f_i(w^*) \leq f_0(w^*)$

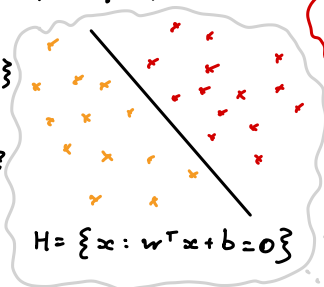
So $\leq \to =$ yields

$$\begin{cases} \sum_{i=1}^m \nu_i^* f_i(w^*) = 0 \Rightarrow \nu_i^* f_i(w^*) = 0 \quad \forall i \in \{1, \dots, m\} & \text{[complementary slackness]} \\ w \in \operatorname{argmin}_w L(w, \xi^*, \nu^*) \Rightarrow \nabla_w L(w^*, \xi^*, \nu^*) = 0 \end{cases}$$

---

(10) **Support Vector Machines**: Data (x_i, y_i) $y_i \in \{-1, 1\}$, $x_i \in \mathbb{R}^p$. Goal: linear classifier $h(x) = \text{sgn}(w^T x + b)$, $w \in \mathbb{R}^p$, $b \in \mathbb{R}$

- How to pick h? Maximize distance to closest data point in both classes. Margin between $H, \{x_i\}$ is $\gamma(w, b) = \min_{1 \leq i \leq n} \{ \|x - x_i\|, i = 1, \dots, n, x \in H \}$
- Distance to hyperplane. Given $x \in \mathbb{R}^p$, $\min_{z \in H} \|x - z\| = ?$

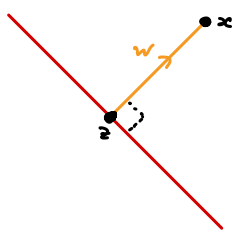


note: $h(x_i) = y_i$ $\Leftrightarrow \text{sgn}(w^T x_i + b) = y_i$ $\Leftrightarrow y_i (w^T x_i + b) > 0$

ie $y_i = -1$, $h(x_i) = -1 \to$ true

$y_i = 1$, $h(x_i) = 1 \to$ true

Ⓐ Geometric argument: w orthogonal to hyperplane so projection of x onto H has form $z = x + tw$ for some $t \in \mathbb{R}$.



Require $z \in H \Rightarrow -b = w^T z = w^T x + t \|w\|^2$

$\Leftrightarrow t = \dfrac{-b - x^T w}{\|w\|^2}$

$\Rightarrow \min_{z \in H} \|x - z\| = |t| \|w\| = \dfrac{|b + x^T w|}{\|w\|}$

So SVM is

$$\max_{w,b} \min_{1 \leq i \leq n} \frac{|b + x_i^T w|}{\|w\|} \quad \text{s.t.} \quad y_i (w^T x_i + b) \geq 0$$

⑥ Dual Approach:

Optimisation problem: $\min_{z \in H} \|x - z\|^2$ s.t. $w^T z + b = 0$ [x is given data]

Lagrange function: $L(z, \xi) = \|x - z\|^2 + \xi(w^T z + b)$

Lagrange Dual function: $D(\xi) = \min_z \|z - x\|^2 + \xi(w^T z + b)$

Strong duality: Holds ∵ convex problem & $\exists z$ s.t. $w^T z + b = 0$

First order optimality: Minimiser in $D(\xi)$ satisfies $\nabla f_0(z_*) + A^T \xi = 0$

so $2(z_* - x) + \xi w = 0 \Rightarrow z_* = -\frac{1}{2}\xi w + x$

Sub min into $D(\xi)$: $D(\xi) = -\frac{1}{4}\xi^2\|w\|^2 + \xi(w^T x + b)$

Dual problem $\max_{\xi \geq 0} D(\xi)$: maximiser ξ^* satisfies $\nabla D(\xi^*) = -\frac{1}{2}\xi^*\|w\|^2 + w^T x + b = 0$

so $\xi^* = \frac{2(b + w^T x)}{\|w\|^2}$ ∴ $z_* = x - \frac{b + w^T x}{\|w\|^2} w$ [First order optimality]

solve w/ optimal z_*: $\|z_* - x\|^2 = \min_{z \in H} \|z - x\|^2 = \frac{|b + w^T x|}{\|w\|}$

Maximum margin classifier for SVM:

Find w, b to max $\gamma(w,b)$

$$\max_{w,b} \min_{1 \leq i \leq n} \frac{|b + w^T X_i|}{\|w\|}$$

s.t. $\forall i \; Y_i(w^T X_i + b) \geq 0$

can assume

- Scale invariance: $\gamma(tw, tb) = \gamma(w, b) \; \forall t \in \mathbb{R} \setminus \{0\} \Rightarrow$ fix $\min_{1 \leq i \leq n} |b + w^T X_i| = 1$

  consider $\operatorname{argmax}_{w,b} \frac{1}{\|w\|}$ s.t. $Y_i(w^T X_i + b) \geq 0$ and $|b + w^T X_i| \geq 1$

- Primal formulation (SVM): Absorb conditions:

$$\operatorname{argmin}_{w,b} \|w\| \quad \text{s.t.} \quad Y_i(w^T X_i + b) \geq 1 \quad \forall i \qquad (P)$$

- Dual Problem SVM: (P) is convex. Assume separating H exists $\Rightarrow$ HAVE STRONG DUALITY! (support vectors: data s.t. $Y_i(w^T X_i + b) = 1$)
  - ↳ Lagrange function: $L(w, b, \xi) = \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \xi_i(1 - Y_i(w^T X_i + b))$ for $\xi \in \mathbb{R}^n_{\geq 0}$, $(w,b) \in \mathbb{R}^{p+1}$
  - ↳ Lagrange Dual: $D(\xi) = \inf_{w,b} L(w,b,\xi) = \inf_w \frac{1}{2}\|w\|^2 + \sum_{i=1}^n \xi_i(1 - Y_i(w^T X_i)) + \inf_b -b \sum_{i=1}^n \xi_i Y_i = -\infty \Rightarrow \langle \xi, Y \rangle = 0$
  - ↳ First order optimality: w, b minimise $D(\xi)$, then

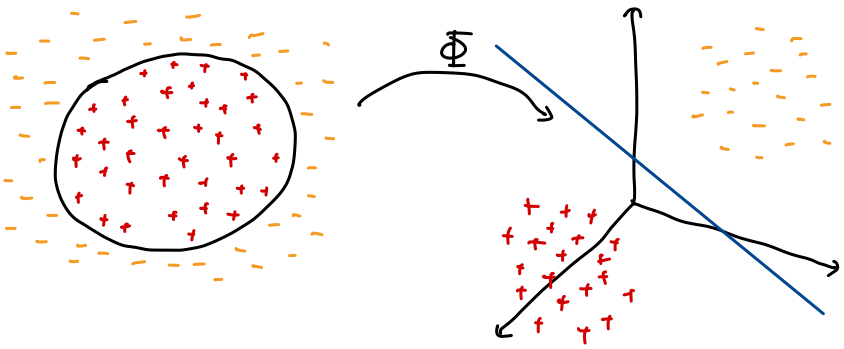
$$\partial_w L(w,b,\xi) = w - \sum_{i=1}^n \xi_i Y_i X_i = 0, \quad \partial_b L(w,b,\xi) = -\sum_{i=1}^n \xi_i Y_i = 0$$

  - ↳ Form of Dual:

$$D(\xi) = \begin{cases} -\frac{1}{2}\|\sum_{i=1}^n \xi_i Y_i X_i\|^2 + \sum_{i=1}^n \xi_i & Y^T \xi = 0 \\ -\infty & \text{o/w} \end{cases}$$

  - ↳ Then use form of dual & complementary slackness: $\xi_i(1 - Y_i(w^T X_i + b)) = 0$
- Non-exact separation: see Q 6.3. You allow some mistakes.
- Non-linear Decision boundaries: Project into higher dim space where you can have a linear separator. Define a feature map $\Phi: X \to \mathbb{R}^d$

$$\Phi(x) = (\varphi_j(x))_{j=1}^d \qquad \varphi_j: X \to \mathbb{R} \text{ basis functions}$$



Hypothesis class:

$$\mathcal{H} = \{ h(x) = w^T \Phi(x) + b : w \in \mathbb{R}^d, b \in \mathbb{R} \}$$

SVM in this setting becomes

$$\operatorname{argmin}_{\substack{w \in \mathbb{R}^d \\ b \in \mathbb{R}}} \|w\| \quad \text{s.t.} \quad Y_i(w^T \Phi(X_i) + b) \geq 1 \qquad (P)$$

$$\max_{\xi \in \mathbb{R}^n} -\frac{1}{2}\left\|\sum_i \xi_i \Phi(X_i) Y_i\right\|^2 + \sum_i \xi_i \qquad (D)$$

$$\text{s.t.} \quad Y^T \xi = 0 \text{ and } \xi \geq 0$$

- Kernels: Goal is to avoid a high dim vector.

Define $k(x, x') = \Phi(x)^T \Phi(x)$

$$(D) \quad \left\|\sum_{i=1}^n \xi_i \Phi(X_i) Y_i\right\|^2 = \sum_{i=1}^n \sum_{j=1}^n \xi_i \xi_j Y_i Y_j k(X_i, X_j)$$

$$= \langle \xi \circ Y, K_n(\xi \circ Y) \rangle$$

$K_n := (k(X_i, X_j))_{i,j=1}^n$ [matrix], $\xi \circ Y = (\xi_i Y_i)_{i=1}^n$ [vector]

Point: dual problem reduces to evaluating the kernel. can avoid handling high dimensional Φ explicitly

↳ Examples:

Ⓐ Polynomial kernels: $k(x, x') = x^T x'$ [linear], $k(x, x') = (x^T x' + 1)^d$ $d \in \mathbb{N}$ [polynomial]

e.g. $p = 2$, $d = 2$, $k(x, z) = (x^T z + 1)^2 = (\sum_{i=1}^2 x_i z_i + 1)^2 = \sum_{i,j=1}^2 (x_i x_j)(z_i z_j) + \sum_{i=1}^2 (\sqrt{2} x_i)(\sqrt{2} z_i) + 1$

so feature vector: $\varphi(x) = (x_1^2 \quad x_1 x_2 \quad x_2 x_1 \quad x_2^2 \quad \sqrt{2} x_1 \quad \sqrt{2} x_2 \quad 1)$

Point: $x \in \mathbb{R}^p$, $k(x, z) = (x^T z + 1)^d \to$ feature space has dimension $\binom{p+d}{d}$ [all monomials → deg $= d$]

while working in $O(p^d)$ dimensional space, computing $k(x, z)$ is $O(p)$

(B) Gaussian kernel: $k(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ $\sigma > 0$

↳ **Practical considerations:**

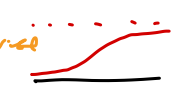
- You choose kernel. Not always obvious (A), (B) popular.
- Choose d in (A) / σ in (B) via cross validation. Big feature map ⇒ high est. error — low approx. error
- kernel SVM works best w/ small/medium sized datasets.

CHAPTER THREE

Neural Networks

set Qs for examples

(11) **Multilayer Perceptrons**: consider functions of the form $h_{w,a,b}(x) = a^T \sigma(Wx+b)$ — applied pointwise



- **Universal Approximation**: Assume $\sigma: \mathbb{R} \to \mathbb{R}$ non-decreasing, cts, $\lim_{r\to-\infty}\sigma(r)=0$, $\lim_{r\to\infty}\sigma(r)=1$

↳ **THM** For any compact set $\Omega \subset \mathbb{R}^p$, the space spanned by the functions [w/ σ as $\uparrow$]

$$\mathcal{F}_\sigma = \left\{ \varphi_{w,b}(x) := \sigma(x^T w + b),\ w \in \mathbb{R}^p,\ b \in \mathbb{R} \right\}$$

is dense in $C(\Omega)$ w/ uniform convergence. This means for ANY continous function $f: \Omega \to \mathbb{R}$ and ANY $\varepsilon > 0$, $\exists q \in \mathbb{N}$, $w_j \in \mathbb{R}^p$, $a_j, b_j \in \mathbb{R}$ s.t.

$$\left| f(x) - \sum_{k=1}^{q} a_k \varphi_{w_k, b_k}(x) \right| \leq \varepsilon$$

THE UNIVERSAL APPROXIMATON THEOREM

Proof: exact proof non-examb. need the stone-weirstrauss & how it can be applied...

[Fund. result in approximation theory]

↳ **The stone weirstrauss thm**: Let $\Omega \subset \mathbb{R}^p$ be compact. Let $\mathcal{F}$ be a sub-algebra of $C(\Omega)$ ie (i) $1 \in \mathcal{F}$, (ii) $f \cdot g \in \mathcal{F}$ (iii) $\alpha f + \beta g \in \mathcal{F}$ $\forall a, b \in \mathbb{R}$ and $\mathcal{F}$ separates points in Ω. ie $\forall x \neq y$ $\exists f \in \mathcal{F}$ s.t. $f(x) \neq f(y)$. Then $\mathcal{F}$ is dense in $C(\Omega)$ w/ uniform norm. ie $\forall g: \Omega \to \mathbb{R}$ cts, $\exists f \in \mathcal{F}$ s.t. $\|f-g\|_\infty \leq \varepsilon$ Pf: Fundamental result 1885

note: point separation necessary. If x, x' cannot be separated ⇒ $\forall f \in \mathcal{F}$ $f(x) = f(x')$ Then define g cts s.t. $g(x) \neq g(x')$ & no f can approx.

E.g. $g(z) = \langle z - x, x' - x \rangle$. $g(x) = 0$, $g(x') = \|x'-x\|^2 \neq 0$

Example: $\forall g \in C(\Omega)$, $\forall \varepsilon > 0$, $\exists f \in \text{Span}\, \mathcal{F}_{exp}$ s.t. $\|f-g\|_\infty \leq \varepsilon$

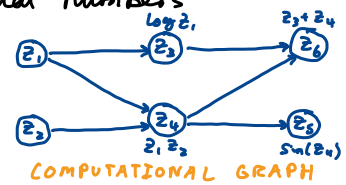
Pf: verify stone-weirstrauss conditions

- clearly all $f \in \mathcal{F}_{exp}$ are continous
- $x \longmapsto \exp(\langle 0, x \rangle) = 1 \in \mathcal{F}_{exp}$
- Span $\mathcal{F}_{exp}$ is a linear subspace, hence an algebra $\because e^{t+s} = e^t e^s$
- Point separation if $x \neq y$, pick $f(z) = \exp(\langle z-x, y-x \rangle / \|y-x\|^2)$
  $f(x) = \exp(0) = 1$, $f(y) = \exp(1) = e$

- **Deep Neural Networks**: Idea is to stack several layers of perceptrons.
  $F^1(x) = \sigma(W^1 x + b^1)$, $F^{k+1} = \sigma(W^{k+1} F^k(x) + b^{k+1})$, $W^k \in \mathbb{R}^{d_k \times d_{k-1}}$, $b^k \in \mathbb{R}^{d_k}$,
  call $F^L(x)$ a neural net of format $d = (d_1, \dots, d_L)$. $\mathcal{H} = \{ F: \mathbb{R}^{d_0} \to \mathbb{R}^{d_L} : F \text{ NN w/ format } d \}$
  ↳ classification: $d_L = k$ x assigned to argmax $F_i(x)$
  ↳ Regression: output $h(x) = a^T F^L(x)$ $a \in \mathbb{R}^{d_L}$

(12) **Automatic Differentiation**: To find params of NN, perform ERM $\min_{h \in \mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} \ell(h(x_i), y_i)$ w/ SGD. Gradient bottleneck.

- Finite difference only gives an approx, Symbolic diff memory intensive. Auto diff is exact. — complexity same as eval. function
- computational graphs
- Forward mode
- Reverse mode
- Dual numbers



COMPUTATIONAL GRAPH

Example: $f(z_1, z_2) = (\sin(z_1 z_2), \log(z_1) + z_1 z_2)$ $\frac{\partial f}{\partial z_k} = \sum_{m \in \text{child}(k)} \frac{\partial f}{\partial z_m} \cdot \frac{\partial z_m}{\partial z_k} = \sum \dot{z}_m$

FORWARD PASS $f(2,3)$

$z_1 = 2$,
$z_2 = 3$
$z_3 = \log 2$
$z_4 = 6$
$z_5 = \sin(6)$
$z_6 = \log 2 + 6$

BACKWARD PASS: Initialize $\dot{z}_6 = 1$, $\dot{z}_5 = 1$

$\dot{z}_4 = \dot{z}_6 \cdot \frac{\partial f_6}{\partial z_4} + \dot{z}_5 \cdot \frac{\partial f_5}{\partial z_4} = 1 \times 1 + 1 \times \cos(z_4) = 1 + \cos(6)$
$\dot{z}_3 = \dot{z}_6 \cdot \frac{\partial f_6}{\partial z_3} = 1 \times 1 = 1$
$\dot{z}_2 = \dot{z}_4 \frac{\partial f_4}{\partial z_2} = (1 + \cos(6))\, z_1 = 2 + 2\cos(6)$
$\dot{z}_1 = \dot{z}_3 \frac{\partial f_3}{\partial z_1} + \dot{z}_4 \frac{\partial f_4}{\partial z_1} = 1 \times \frac{1}{z_1} + (1+\cos(6)) \times z_2 = \frac{1}{2} + 3 + 3\cos(6)$

So $\nabla f(2,3) \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 3\cos 6 & 2\cos 6 \\ \frac{1}{2}+3 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 5\cos 6 \\ 5 + \frac{1}{2} \end{pmatrix}$

↳ Gradient descent is a special case of this w/ $f: \mathbb{R}^p \to \mathbb{R}$.

(discrete)

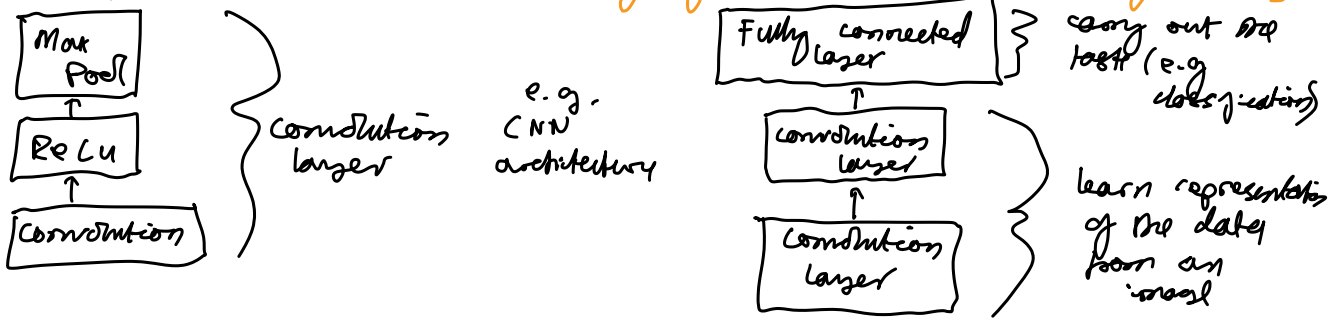
(13) **Cross-Entropy Loss:** Two prob. distributions P, Q on same sample space Ω.

- KL-divergence between P & Q is $KL(P|Q) = \sum_{\omega \in \Omega} \log\left(\frac{P(\omega)}{Q(\omega)}\right) P(\omega)$
  - ↳ $Q(\omega) = 0 \Rightarrow P(\omega) = 0$
  - ↳ $KL(P|Q) \geq 0$ and $KL(P|Q) = 0$ if $P = Q$
- If class classification: $F(x) = (v_1, \dots, v_K)$
  - ↳ To output a probability vector, post process via $p_k = \frac{e^{v_k}}{\sum_{j=1}^{K} e^{v_j}}$
  - ↳ $\ell(F(x), y) = -\sum_{j=1}^{K} y_j \log(p_j)$ [cross entropy loss]

Note: Binary classification $\ell(F(x), y) = -y \log p - (1-y)\log(1-p)$, $p = \frac{1}{1+e^{-v}}$
minimizes $\sum \log(1 + e^{-(y_i w^T x_i)})$

(14) **Convolutional Neural Networks:** use in image processing applications

- Convolutions / Fourier transforms
  - ↳ $(f * g)(t) = \int f(x) g(t-x)\, dx$
  - ↳ won't cover up surely...
- CNN: a feedforward NN $x^{k+1} = \sigma(W^k x^k + b^k)$ $x^k \in \mathbb{R}^{n_k \times d_k}$
  where n_k = # spatial positions, d_k = # channels e.g. RGB $\Rightarrow d_k = 3$.
  - ↳ spatial position ↑ & to ensure translation invariance, linear operator is a convolution operator. [convolutions are the only linear operators that are translation invariant]
  - ↳ max pool to ↓ # pixels. [summary of info in a window of size r]



(15) **Generative Adversarial Networks:** Goal: learn the distribution ρ_X of the input X

- GANS. You have
  - Data space X w/ pdf ρ_X
  - Latent space Z w/ pdf ρ_Z [compressed rep. of ρ_X]
  - $G: Z \to X$ generator
  - $D: X \to [0,1]$ discriminator.

GAN is a (D, G) pair

GOALS:

- Discriminator D aims to discriminate data sampled from ρ_X and ρ_Z.
- Generator G aims to generate data in X that is indistinguishable from ρ_X.

↳ Pushforward measure $\rho_G = G_* \rho_Z$

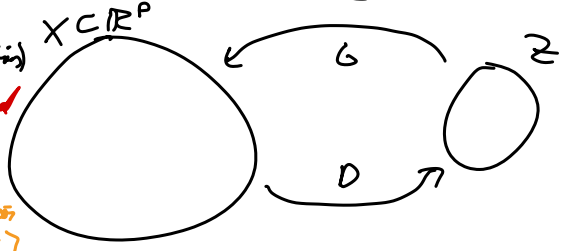
$$\int_X f(x) \rho_G(x)\, dx = \int_Z f(G(z)) \rho_Z(z)\, dz \qquad \forall f: X \to \mathbb{R} \text{ measurable.}$$

Aim: $\rho_G = \rho_X$.

$$\min_G \max_D V(G,D), \quad V(G,D) = \int \log(D(x)) \rho_X(x)\, dx + \int \log(1 - D(x)) \rho_G(x)\, dx$$

(16) Variational Auto-Encoders [compact representation of high dimensional data]

- why? To visualise data, ↓ noise, ↓ resources, ↑ performance of ↓ stream tasks (e.g. classification)
- Idea: Encoder $E: X \longrightarrow Z$
  $\mathbb{R}^p \ni x \longmapsto G(x) = z \in \mathbb{R}^d$ [lower dimensional representation]
  Decoder $D: Z \longrightarrow X$
  $z \longmapsto D(z)$ [Takes a representation & reconstructs it]



- Optimisation problem: $p >> d$

$$\min_{E,D} \frac{1}{n} \sum_{i=1}^{n} \| D(E(x_i)) - x_i \|^2 \quad \text{w/ data } \{x_i\}_{i=1}^n \subset \mathbb{R}^p$$

- VAE: model D & E as NNs,
  ↳ $z \sim q_\theta(\cdot|x)$ ↳ GIVEN Data
  ↳ $x \sim p_\theta(\cdot|z)$ ↳ representation [sample from these conditional probability distributions]
  To generate data, new sample $z \sim p_z$ then draw from data space $p_\theta(\cdot|z)$
  ↳ Model the joint distribution as $p(x,z) = p_\theta(x|z)\, p_z(z)$
  ↳ Pick $p_z = \mathcal{N}(0, I_d)$
  ↳ model $p_\theta(\cdot|z) = \mathcal{N}(G_\theta(z), \sigma_\theta^2 I_n)$ where $G_\theta: \mathbb{R}^d \longrightarrow \mathbb{R}^p$ [neural network]
  This is good: p_z & $p(\cdot|z)$ simple gaussians & can model complex p_{data}

$$p_\theta = \int p_\theta(x|z)\, p_z(z)\, dz$$

Useful Formulae

① $\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|W]]$ [Tower property]

② $\mathbb{E}[g(W)Z|W] = g(W)\mathbb{E}[Z|W]$ [Taking out known]

② $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$ [Bayes Theorem] Pf: $\mathbb{P}(A|B) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(B)}$, $\mathbb{P}(B|A) = \frac{\mathbb{P}(A\cap B)}{\mathbb{P}(A)}$

③ $\inf_{\|w\|\le k}\{\|w - w_*\|\} = \max\{\|w_*\| - k, 0\}$ [i.s.i approximation error e.g.]

④ p is a descent direction of f if $\langle p, \nabla f\rangle < 0$

⑤ $f(x+\varepsilon) = f(x) + \langle \nabla f(x), \varepsilon\rangle + h.o.t.$ [Taylor expansion]

⑥ $\sigma(at) = \frac{e^{at}}{1+e^{at}} \xrightarrow{a\to\infty} \mathbb{1}_{[0,\infty)}$ [Sigmoid is cts approx to heaviside]