Von Neumann's alg VA:

1. Sample $X \sim \mathrm{Bern}(p)$
   $Y \sim \mathrm{Bern}(p)$
2. If $X \neq Y$:
   return X
3. If $X = Y$:
   return VA()

Goal: given access to a biased coin $\mathrm{Bern}(p)$ we want to simulate a draw from a fair coin:

$$\mathbb{P}(VA() = 1) = \frac{1}{2}$$

Q: what is the expected runtime (measured in coin flip) of this algorithm?

$$T = \begin{cases} 2 & X \neq Y \\ 2 + T' & X = Y \end{cases}$$

where $T' \sim T$
and $T' \perp\!\!\!\perp T$
and $T' \perp\!\!\!\perp (X \,\&\, Y)$

if T & ω are r.v.s. then $\mathbb{E}[T] = \mathbb{E}_{\omega}[\mathbb{E}[T|\omega]]$

tower property

To apply tower property, take

$$\omega = \begin{cases} 1 & \text{if } X \neq Y \\ 0 & \text{if } X = Y \end{cases}$$

note ω is a function of X & Y, so since $T' \perp\!\!\!\perp (X \& Y)$, ω is also indep of T'

$\mathbb{E}[T|\omega=1] = \mathbb{E}[2|\omega=1]$
$= 2$

Use the tower property

$$\mathbb{E}[T] = \mathbb{E}_{W}[\mathbb{E}[T|\omega]] = \mathbb{E}[T|\omega=0]\mathbb{P}(\omega=0) + \underline{\mathbb{E}[T|\omega=1]}\mathbb{P}(\omega=1)$$

$$= \mathbb{E}[2+T'|\omega=0]\mathbb{P}(\omega=0) + 2\,\mathbb{P}(\omega=1)$$

$$\mathbb{E}[T] = 2\,\underline{\mathbb{P}(W=1)} + \underbrace{\mathbb{E}[2+T' \mid W=0)}\,\underline{\mathbb{P}(W=0)}$$

and $\mathbb{P}(W=1) = \mathbb{P}(X \neq Y) = \mathbb{P}(\{X=H, Y=T\} \cup \{X=T, Y=H\})$

$$= \mathbb{P}(X=H, Y=T) + \mathbb{P}(X=T, Y=H)$$

$$= \mathbb{P}(X=H)\mathbb{P}(Y=T) + \mathbb{P}(X=T)\mathbb{P}(Y=H)$$

$$= p(1-p) + (1-p)p = 2p(1-p)$$

$$\Rightarrow \mathbb{P}(W=0) = 1 - \mathbb{P}(W=1)$$

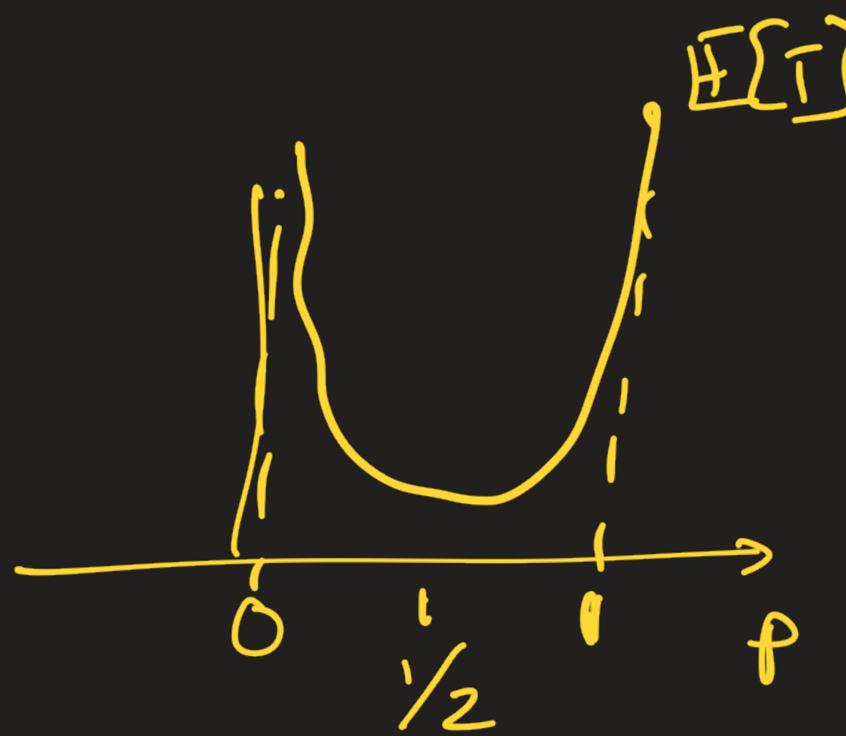
$$= 1 - 2p(1-p)$$

$$\mathbb{E}[2+T' \mid \omega = 0] = 2 + \mathbb{E}[T' \mid \omega = 0]$$

$$T' \perp\!\!\!\perp \omega \Rightarrow \mathbb{E}[T' \mid \omega] = \mathbb{E}[T']$$

$$= 2 + \mathbb{E}[T']$$

$$= 2 + \mathbb{E}[T]$$



Altogether

$$\mathbb{E}[T] = \underline{2 \cdot 2p(1-p)} + (\underline{2} + \mathbb{E}T)\,\underline{(1-2p(1-p))}$$

$$= 2 + \mathbb{E}T\,(1-2p(1-p))$$

$$\Rightarrow \mathbb{E}[T]\,\underbrace{(1-[1-2p(1-p)])}_{=\,2p(1-p)} = 2 \Rightarrow \boxed{\mathbb{E}[T] = \frac{1}{p(1-p)}}$$

Transformations of random variables

Begin with, $y = f(x)$ ($x, y \in \mathbb{R}$ are r.v.s)

and f be invertible & monotonic & continuous

monotonic increasing

Question: give $p_x(x)$

determine $p_y(y)$

monotonic $\Rightarrow$ if $x > y$ then

increasing $f(x) \geq f(y)$

or

decreasing $f(x) \leq f(y)$

Recall def of the density function implies

$$\int_y^{y+\Delta y} p_y(y) = \mathbb{P}(y \in [y, y+\Delta y])$$

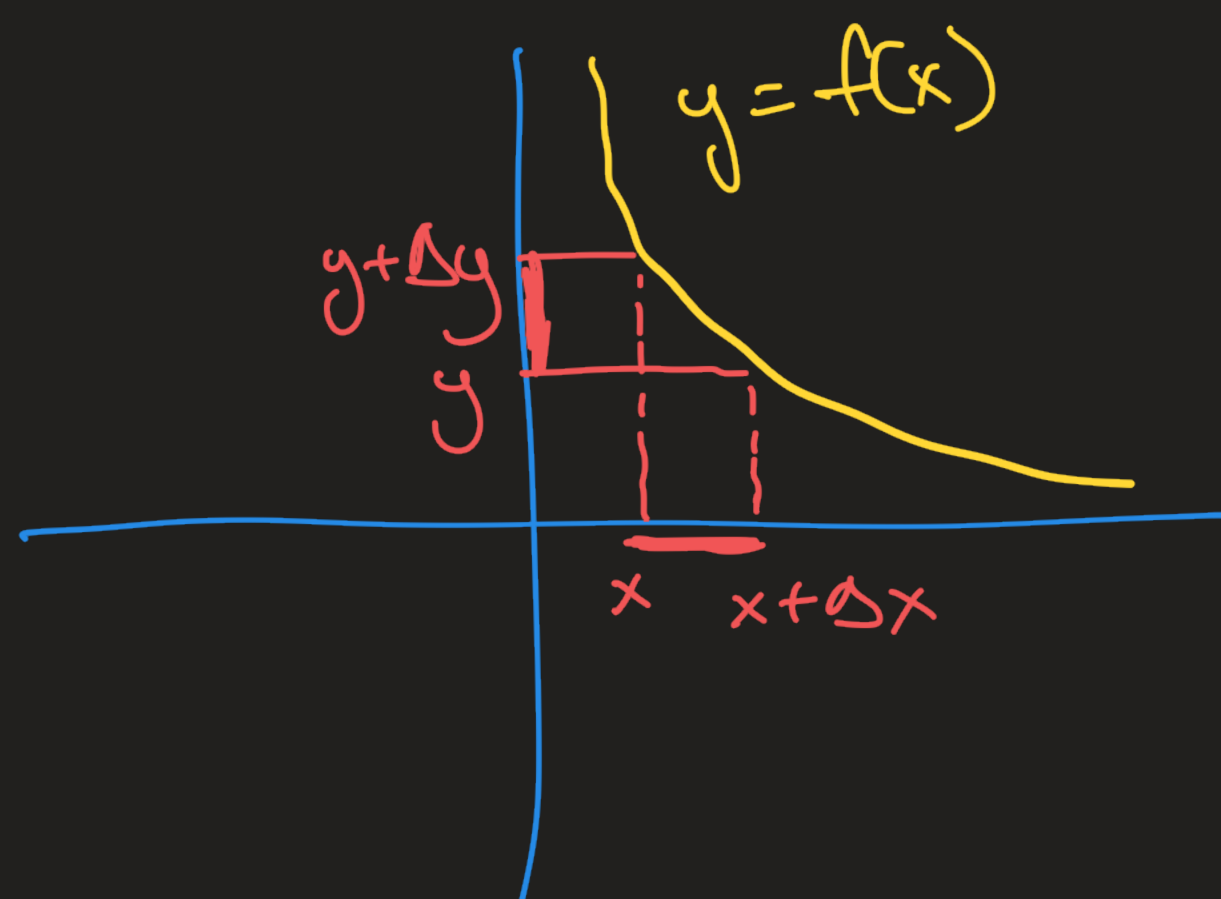
and

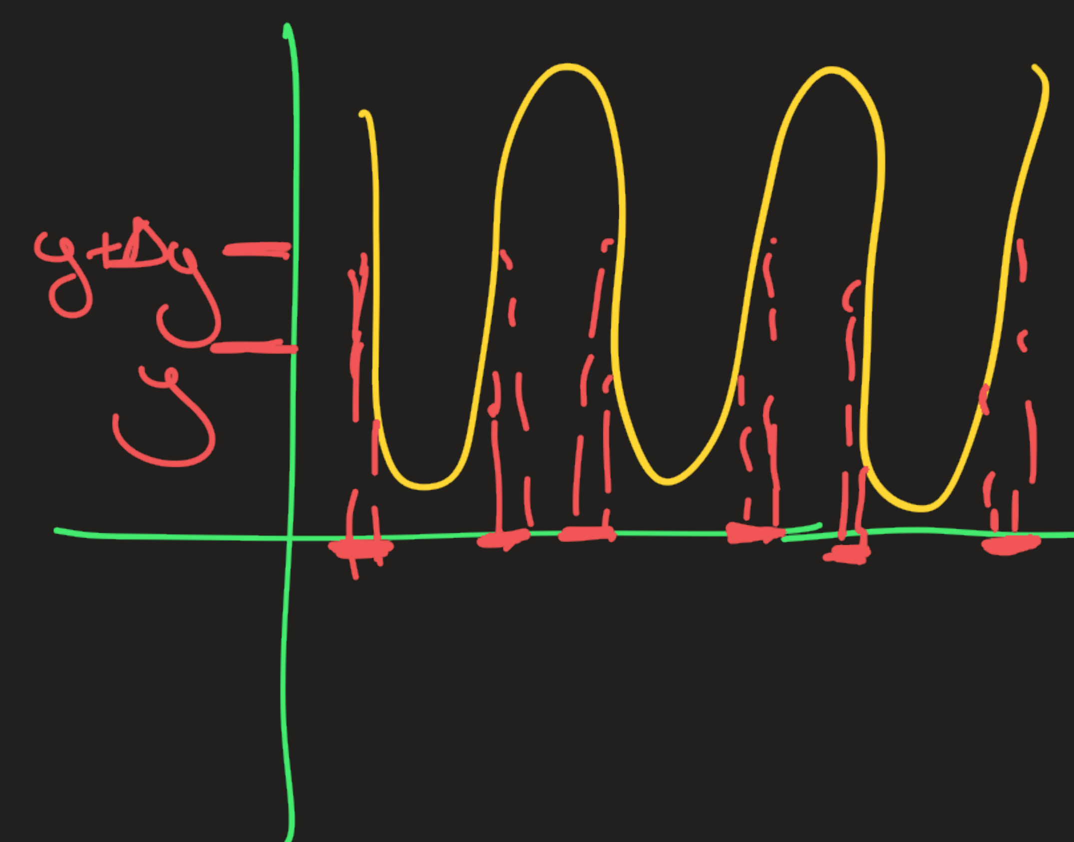
$$\int_x^{x+\Delta x} p_x(x) = \mathbb{P}(x \in [x, x+\Delta x])$$

Since f is invertible and monotonic,

$$\mathbb{P}(y \in [y, y+\Delta y]) = \mathbb{P}(f(x) \in [y, y+\Delta y])$$

$$= \mathbb{P}(x \in [f^{-1}(y), f^{-1}(y+\Delta y)])$$





$$\approx \mathbb{P}(x \in [f^{-1}(y), f^{-1}(y) + \Delta x])$$

since $y=f(x)$

$$= \mathbb{P}(x \in [x, x+\Delta x])$$

$$\Rightarrow \int_{y}^{y+\Delta y} p_y(y)\,dy \approx \int_{x}^{x+\Delta x} p_x(x)\,dx$$

$$\wr\wr$$

$$p_y(y)\,|\Delta y| \approx p_x(x)\,|\Delta x|$$

$$\Rightarrow \text{let } \Delta y \to 0 \qquad \boxed{p_y(y)\left|\frac{dy}{dx}\right| = p_x(x)}$$

$$\Rightarrow p_y(y) = \frac{p_x(x)}{\left|\frac{dy}{dx}\right|} = \frac{p_x(f^{-1}(y))}{\left|\frac{dy}{dx}(f^{-1}(y))\right|}$$

For the multivariable case, if $y \in \mathbb{R}^k$ and $x \in \mathbb{R}^d$
and $y = f(x)$

f invertible

$$J_y = \left[\frac{\partial y_i}{\partial x_j} \right]_{ij} \in \mathbb{R}^{k \times d}$$

$$= \begin{bmatrix} \frac{\partial f(x_1, \dots, x_d)_1}{\partial x_1} & \cdots & \frac{\partial f(x_1, \dots, x_d)_1}{\partial x_d} \\ \vdots & & \vdots \\ \frac{\partial f(x_1, \dots, x_d)_k}{\partial x_1} & \cdots & \frac{\partial f(x_1, \dots, x_d)_k}{\partial x_d} \end{bmatrix}$$

$$p_y(y) = \frac{p_x(x)}{|\det J_y|} = \frac{p_x(f^{-1}(y))}{|\det J_y(f^{-1}(y))|}$$

<u>Ex</u> $Y = e^X$, $X \sim \mathcal{N}(\mu, \sigma^2)$ Y is log-normal

Then $y = f(x)$ where $f(x) = e^x$

$$p_y(y) = \frac{p_x(f^{-1}(y))}{|f'(f^{-1}(y))|} = \frac{p_x(\ln y)}{|e^{\ln y}|} = \frac{p_x(\ln y)}{|y|}$$

$$= \frac{1}{y} p_x(\ln y)$$

$$= \frac{1}{y\sqrt{2\pi\sigma^2}} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}}$$

Multivariate Gaussians

- completely determined by μ, Σ, $\mathcal{N}(\mu, \Sigma)$
- $p_x(x) = \dfrac{1}{|\det(2\pi\Sigma)|^{1/2}} \exp\left[-\dfrac{(x-\mu)^T \Sigma^{-1}(x-\mu)}{2}\right]$

facts:

1) sums of Gaussians are Gaussians

if $X \sim \mathcal{N}(\mu_1, \Sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \Sigma_2)$, $X \perp\!\!\!\perp Y$

then

$$cX + Y \sim \mathcal{N}(c\mu_1 + \mu_2,\ c^2\Sigma_1 + \Sigma_2)$$

where c is a scalar

2) if U is orthonormal $(UU^T = I = U^TU)$ and $X \sim \mathcal{N}(0, I)$

then $UX \sim \mathcal{N}(0, I)$

Interpret Gaussians

- start off w/ standard Gaussians: $\mu = 0$ and $\Sigma = I_d$

$$p_x(x) = \frac{1}{|\det(2\pi I)|^{1/2}} \exp\left(\frac{-x^T I x}{2}\right)$$

$$= \frac{1}{|(2\pi)^d|^{1/2}} \exp\left(\frac{-\|x\|_2^2}{2}\right)$$

$$\exp\left(\frac{-\sum_{i=1}^{d} x_i^2}{2}\right) = \prod_{i=1}^{d} \exp\left(\frac{-x_i^2}{2}\right)$$

$$= \prod_{i=1}^{d} \left[\frac{1}{\sqrt{2\pi}} \exp\left(\frac{-x_i^2}{2}\right)\right]$$

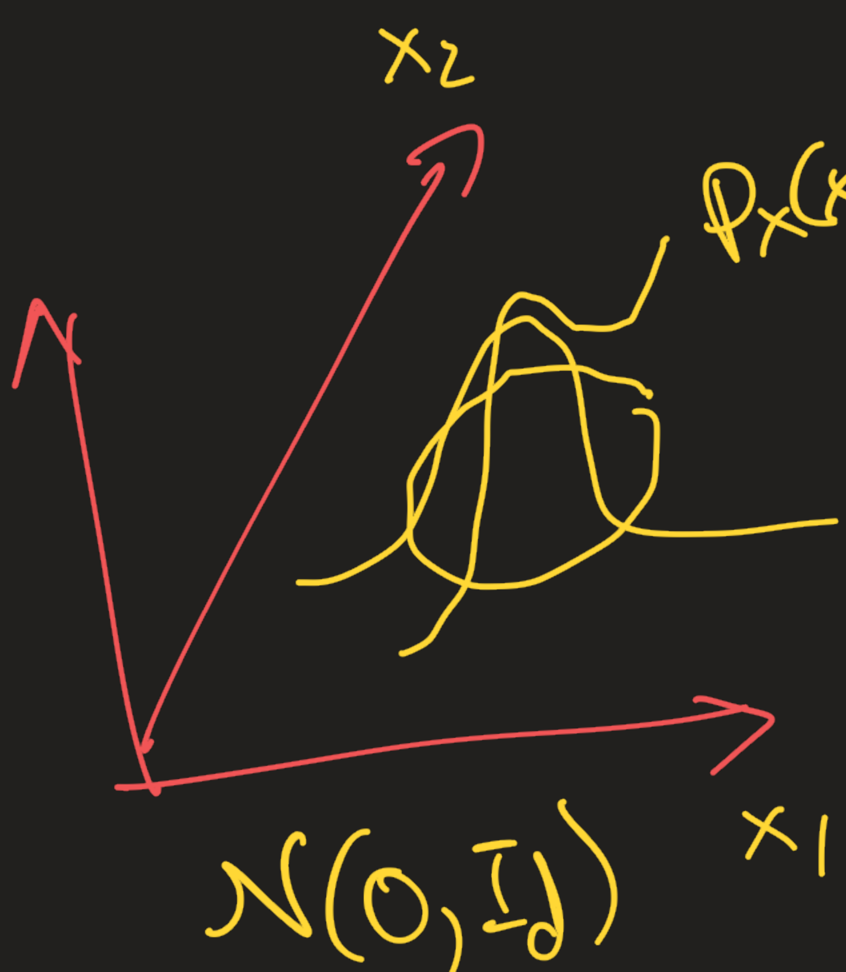
$\Rightarrow$ $x \sim \mathcal{N}(0, I)$ the coordinates are indep, identical $\mathcal{N}(0,1)$

Recall $X \perp\!\!\!\perp Y$ iff $p(x,y) = p_X(x)\,p_Y(y)$

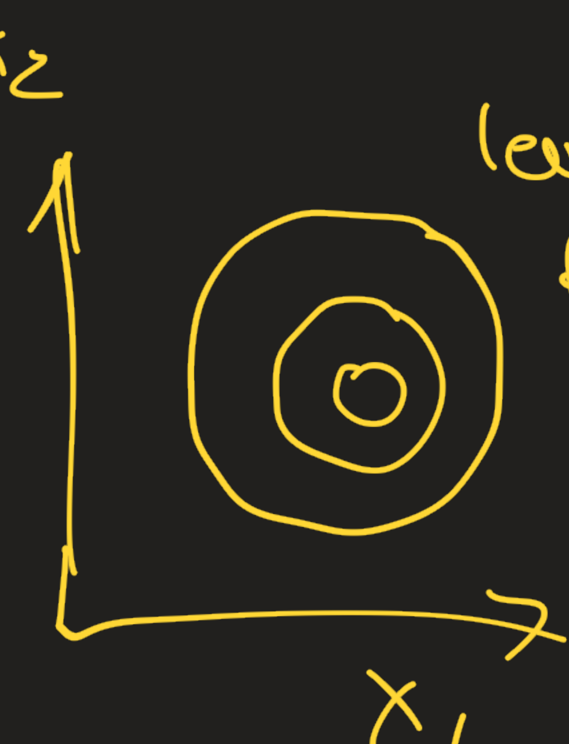
X, Y, Z are indep iff $p(x,y,z) = p_X(x)\,p_Y(y)\,p_Z(z)$

General mult. Gaussian

$$x \sim \mathcal{N}(\mu, \Sigma) \quad \text{where} \quad p_X(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right)$$

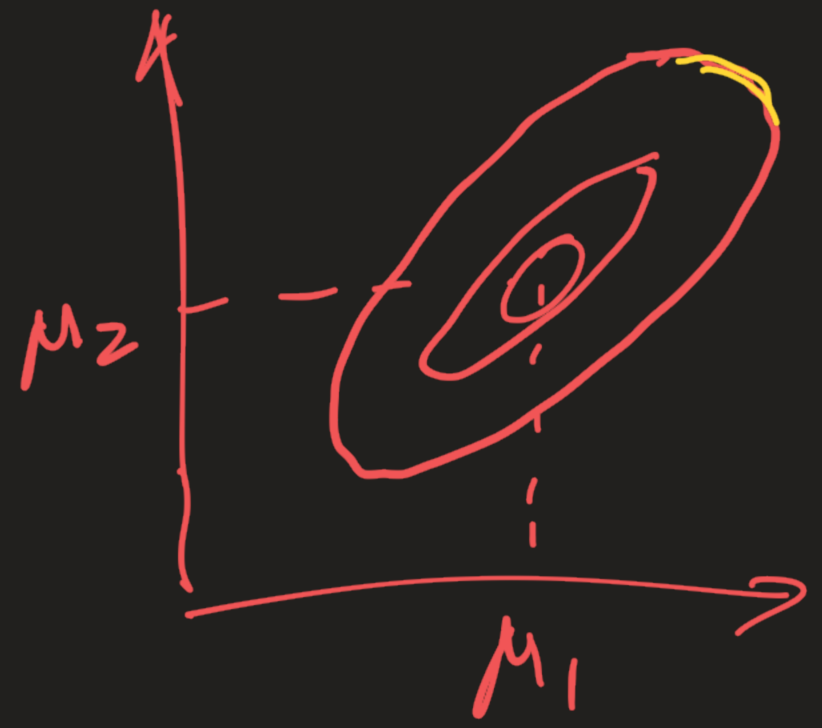


$\longrightarrow$



level sets of $p_X(x)$ for $\mathcal{N}(0, I_d)$

$\mathcal{N}(\mu, \Sigma)$



if $y \sim \mathcal{N}(0, I_d)$

and $\Sigma = U \Lambda U^T$ eigendecomposition

where $U U^T = U^T U = I$

$\Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$

where $\lambda_i > 0$

if

$x = U \Lambda^{1/2} y + \mu$ then $x \sim \mathcal{N}(\mu, \Sigma)$

(put on HW)

Machine Learning as Risk Minimization

We can ~~cast~~ ML as estimate? y from X on some population

$$(X, y) \sim \mathbb{P}$$

could be:

$$p(y \mid X) \leftarrow \text{density estimation}$$

$$\mathbb{E}[y \mid X] \leftarrow \text{regression}$$, e.g. given age, race, sex, smoker, diabetes, compute the expected time until death

$$c^* = \underset{c \in \{1, \dots, K\}}{\operatorname{argmax}} \; p(y = c \mid X) \leftarrow \text{classification}$$

$\vdots$ others (quantile est., ...)

We can use the empirical risk minimization framework to pose many of these questions as optimization problems.

In risk minimization we want to fit a model/hypothesis/function f that minimizes some risk over our population

$$f^* = \underset{f}{\operatorname{argmin}}\ R(f)$$

where

$$R(f) = \mathbb{E}_{(x,y)}\, \ell(f(x), y)$$

risk: average loss over our population

loss

Examples of losses

if $y \in \{-1,+1\}$ can use $\ell(u,v) = \begin{cases} 1 & \text{if } u \neq v \\ 0 & \text{if } u = v \end{cases}$

"zero-one loss"

$$R(f) = \mathbb{E}_{(x,y)} \underline{\ell(f(x), y)}$$

$\to$ Bern r.v. $= \begin{cases} 1 & \text{when } f(x) \neq y \\ 0 & \text{when } f(x) = y \end{cases}$

$$= \mathbb{P}(f(x) \neq y)$$

$$f^* = \operatorname*{argmin}_f R(f) = \operatorname*{argmin}_f \mathbb{P}(f(x) \neq y)$$

$\hookrightarrow$ misclassification rate

called the Bayes optimal estimator

Optimizing this loss is intractable, so we use <u>surrogate losses</u>

"hinge-loss" $\ell(u,v) = (1-uv)_+ = \max\{1-uv, 0\}$

"squared hinge-loss" $\ell(u,v) = (1-uv)_+^2$

$\vdots$

Problem: we cannot minimize over all functions

Instead we minimize over a class of function/ hypothesis space $\mathcal{F}$

$$\tilde{\mathcal{F}}_{lin} = \{ f(x) : f(x) = w^T x + b \ \text{ for some } w, b \}$$

$$\mathcal{F}_{neuralnets} = \{ f(x) : \sigma(W_L(\sigma(W_{L-1}(\cdots \sigma(W_1 x + b_1) \cdots) + b_\ell) \ \text{ for some } W_L, W_{L-1}, \dots, W_1 \text{ and } b_1, \dots, b_\ell \}$$

and we define $$f^*_{\tilde{\mathcal{F}}} = \operatorname*{argmin}_{f \in \mathcal{F}} R(f)$$

Another problem, we can't compute $R(f)$

b/c

$$R(f) = \mathbb{E}_{\underbrace{(x,y)\sim\mathbb{P}}_{\text{UNKNOWN}}}\, \ell(f(x), y)$$

Define the empirical risk corresponding to n i.i.d training samples $(x_i, y_i) \sim \mathbb{P}$

$$R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$$

Note that for each (x_i, y_i) we have

$$\mathbb{E}\, \ell(f(x_i), y_i) = R(f)$$

so by LLN,

$$R_n(f) = \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)}_{} \approx R(f)$$

and thus we expect

empirical risk minimizer

$$\hat{f}_{\mathcal{F}} = \operatorname*{argmin}_{f \in \mathcal{F}} R_n(f)$$

satisfies

$$R(\hat{f}_{\mathcal{F}}) \approx R(f^*_{\mathcal{F}})$$

Ex linear regression & logistic regression