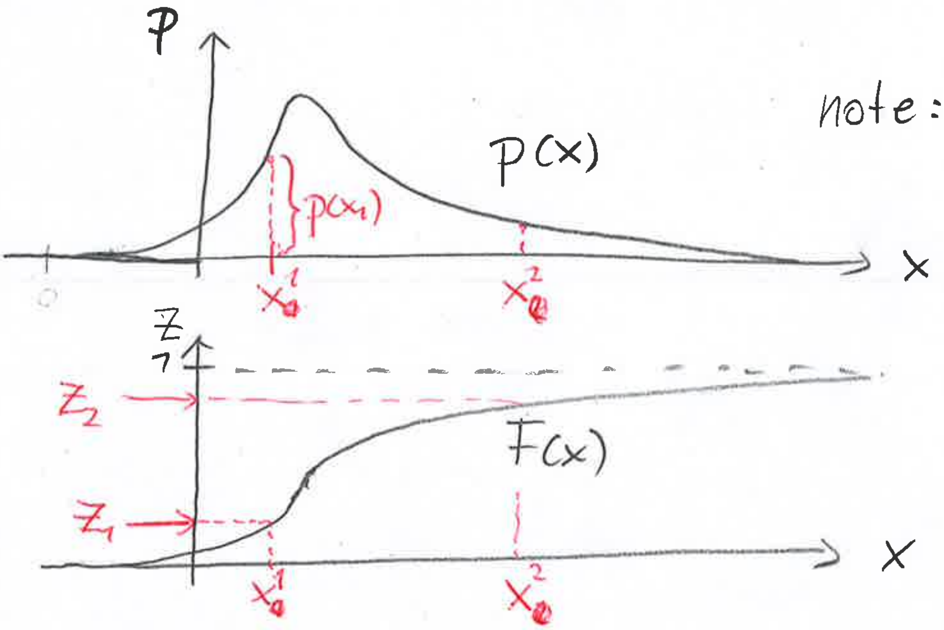
# Blackboard 3.1 : Random generation of data

You want to generate data points

$$\mathcal{X} = \{x^1, x^2, x^3, x^4\}$$

from a distribution $p(x)$




note: $\int p(x)\,dx = 1$

## Q1: How do you do this on a computer?

- Integrate

  define $F(x) = \int_{-\infty}^{x} p(x')\,dx'$

- Draw random numbers $z_k \in [0,1]$
- $x^k = F^{-1}(z_k)$

## Q2: What is the "likelihood" that you generate a point $x^1$?

$P(x^1)$ $\boxed{\text{note: Prob} = p(x)\cdot\Delta x}$

and all point $x^1, x^2, x^3, x^4$

$$P(\mathcal{X}) = P(x^1) \cdot p(x^2) \cdot p(x^3) \cdot p(x^4)$$

independence $(\Delta x)^4$

# Blackboard 3.2: ML for Gaussian

A direct calculation

$$P_{model}(\mathcal{X}|x^{center}) = p(x^1)\cdot p(x^2)\cdot \ \ldots \ p(x^P)$$

$$= \prod_k \left[\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left\{-\frac{(x^k - x_{center})^2}{2\sigma^2}\right\}\right]$$

$$= \left[\frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\right]^K \exp\left\{-\frac{1}{2\sigma^2}\sum_k (x^k - x_{center})^2\right\}$$

optimize parameter $x_{center}$:

$$0 \stackrel{!}{=} \frac{\partial}{\partial x_{center}} P(\mathcal{X}|x^{center})$$

$$= \underbrace{P_{model}(\mathcal{X}|x^{center})}_{\neq 0}\cdot(-1)\frac{1}{\cancel{2}\sigma^2}\cdot\cancel{2}\cdot\sum_k (x^k - x_{center})$$

$$\Rightarrow \boxed{x_{center} = \frac{1}{P}\sum_{k=1}^{P} x^k}$$

B: alternatively with log-likelihood

$$\mathcal{L}(\mathcal{X}|x^{center}) = \ln P_{model}(\cdot|\cdot) = \sum_k \Big[\underbrace{\ln\frac{1}{\sqrt{2\pi}} + \ln\frac{1}{\sigma}}_{\text{constant}} - \frac{(x^k - x_{center})^2}{2\sigma^2}\Big]$$

# Blackboard 3.3A : stochastic model

network output

(1a) $\hat{y}_{\vec{w}}(\vec{x}) = g^{(2)}\left[\sum_i \underline{w_i^{(2)}}\, g^{(1)}\left(\sum_k \underline{w_{ik}^{(1)}}\, x_k\right)\right]$

generation of labels

(1b) $\hat{y}_{\vec{w}} = p(z=+1\,|\,\vec{x})$

label generated by my model

---

**A** study point $\vec{x}^{\mu}$ with $t^{\mu} = +1$

label in data base

What is the probability that $(\vec{x}^{\mu}, +1)$ could have been generated by (1a), (1b)?

$$p(\vec{x}^{\mu}, z^{\mu}=+1) = p(z^{\mu}=+1\,|\,\vec{x}^{\mu}) \cdot p(\vec{x}^{\mu})$$
$$= \hat{y}_{\vec{w}}(\vec{x}^{\mu}) \cdot p(\vec{x}^{\mu})$$

**B** study point $\vec{x}^{\mu}$ with $t^{\mu} = 0$

label in data base

$$p(\vec{x}^{\mu}, z^{\mu}=0) = p(z^{\mu}=0\,|\,\vec{x}^{\mu}) \cdot p(\vec{x}^{\mu})$$
$$= \left(1 - \hat{y}_{\vec{w}}(\vec{x}^{\mu})\right) \cdot p(\vec{x}^{\mu})$$

# Blackboard 3.3 B (continued)

likelihood that set of all points

$$\mathcal{X} = \{ (\vec{x}^\mu, t^\mu) \ ; \ 1 \le \mu \le P \}$$

could have been generated by model

$$P(\mathcal{X}) = \Big[ \prod_{\substack{\vec{x}^\mu \in C \\ \Leftrightarrow t^\mu = +1}} \big( \hat{y}_{\vec{\omega}}(\vec{x}^\mu) \big) \Big] \cdot \Big[ \prod_{\substack{\vec{x}^\mu \notin C \\ \Leftrightarrow t^\mu = 0}} \big( 1 - \hat{y}_{\vec{\omega}}(\vec{x}^\mu) \big) \Big] \cdot \Big[ \prod_\mu p(\vec{x}^\mu) \Big]$$

$$= \prod_\mu \Big[ \big( \hat{y}_{\vec{\omega}}(\vec{x}^\mu) \big)^{t^\mu} \cdot \big( 1 - \hat{y}_{\vec{\omega}}(\vec{x}^\mu) \big)^{(1-t^\mu)} \Big] \cdot \Big[ \prod_\mu p(\vec{x}^\mu) \Big]$$

log-likelihood

$$E(\vec{\omega}) = -\ln P(\mathcal{X}) = -LL_{\vec{\omega}}$$

$$= -\sum_{\mu=1}^{P} \Big[ t^\mu \cdot \ln \big( \hat{y}_{\vec{\omega}}(\vec{x}^\mu) \big) + (1 - t^\mu) \ln \big( 1 - \hat{y}_{\vec{\omega}}(\vec{x}^\mu) \big) \Big]$$

$$\cancel{- \sum_\mu \ln p(\vec{x}^\mu)}$$

constant, does not depend on $\vec{\omega}$

$E(\vec{\omega})$ : cross-entropy error function

minimize with respect to parameters $\vec{\omega}$

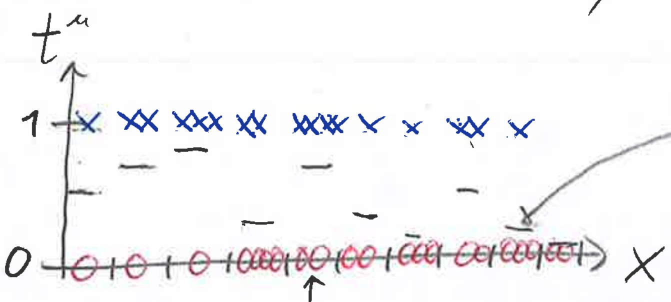
# Blackboard 3.4 : output = probability?

start with

$$E = -\sum_{\mu=1}^{P}\left[t^\mu \cdot \ln \hat{y}^\mu + (1-t^\mu)\ln(1-\hat{y}^\mu)\right]$$

$$= -\sum_{\vec{x}^\mu \in C} \ln \hat{y}^\mu - \sum_{\vec{x}^\mu \notin C} \ln(1-\hat{y}^\mu)$$

hypothesis A : we have many examples




best output

$$\hat{y}(x_k) = \frac{n_1(x_k)}{n_0(x_k)+n_1(x_k)} = p(C \mid x_k)$$

bin $x_m$

# of positive examples in $x_m$

$$n_1(x_m) = p(x, C)\cdot \Delta x$$

# of neg. examples in $x_m$

$$n_0(x_m) = p(x, \bar{C})\cdot \Delta x$$

not in class

bin size

rewrite

$$E = -\sum_{\substack{\text{bins}\\ m}}\left[n_1(x_m)\cdot \ln \hat{y}(x_m) + n_0(x_m)\cdot \ln\left(1-\hat{y}(x_m)\right)\right]$$

hypothesis B : network is flexible enough

$\Rightarrow$ in each bin, $\hat{y}(x_m)$ is arbitrary $\Rightarrow$ optimize!

$$0 = \frac{\partial E}{\partial \hat{y}(x_m)} = \frac{n_1(x_m)}{\hat{y}(x_m)} - \frac{n_0(x_m)}{1-\hat{y}(x_m)}$$

$$\Rightarrow 0 = n_1(x_m)\cdot(1-\hat{y}(x_m)) - \hat{y}(x_m)\cdot n_0(x_m) \Rightarrow \boxed{\hat{y}(x_m) = \frac{n_1(x_m)}{n_0(x_m)+n_1(x_m)}}$$

# Blackboard 3.5: sigmoidal

output as probability

$$\hat{y}_1 = p(C_1|x) \overset{\text{Bayes}}{=} \frac{p(\vec{x}|C_1)\cdot p(C_1)}{p(x)}$$

$$= \frac{p(x,C_1)}{p(x)} = \frac{p(x,C_1)}{p(x,C_1)+p(x,\overline{C_1})}$$

$$= \frac{1}{1+\frac{p(x,\overline{C_1})}{p(x,C_1)}} = \frac{1}{1+b}$$

$0 < b < \infty$

positive parameter, harder to treat analytically

$$= \frac{1}{1+e^{-a}}$$

$$a = \ln\left[\frac{p(\vec{x},C_1)}{p(\vec{x},\overline{C_1})}\right]$$

$a =$ "log-probabilityratio"

$-\infty < a < \infty$

unconstrained parameter

# Blackboard 3.6: mutually exclusive classes

example: 4 symbols A, B, C, D

| with prob | symbol | 1-hot-code |
|---|---|---|
| $P_A$ | A | $= \{1, 0, 0, 0\}$ |
| $P_B$ | B | $= \{0, 1, 0, 0\}$ |
| $P_C$ | C | $= \{0, 0, 1, 0\}$ |
| $P_D$ | D | $= \{0, 0, 0, 1\}$ |
| arbitrary | $\vec{t}$ | $= \{t_1, t_2, t_3, t_4\}$ "1-hot-coding" |

probability to gen. arbitrary symbol $\vec{t}$

$$P_{\vec{t}} = P_A^{t_1} \cdot P_B^{t_2} \cdot P_C^{t_3} \cdot P_D^{t_4} = \prod_i [P_i]^{t_i}$$

check for symbol "C"

total probability to generate M observed target vectors

$$P^{tot} = \prod_{\mu=1}^{M} \prod_i [P_i^\mu]^{t_i^\mu}$$

all patterns ($\mu$), all outputs ($i$)

neg. log-likelihood

$$\boxed{E = -LL = -\ln P^{tot} = -\sum_{\mu=1}^{M} t_i^\mu \ln [P_i^\mu]}$$

probabilities $\sum_i P_i^\mu = 1$

$\Longrightarrow$ describe $P_i^\mu$ by softmax!

$$\boxed{y_i^\mu = \frac{e^{a_i}}{\sum_k e^{a_k}}}$$