Random features and kernels

Goals: law of large numbers
random features as kernels

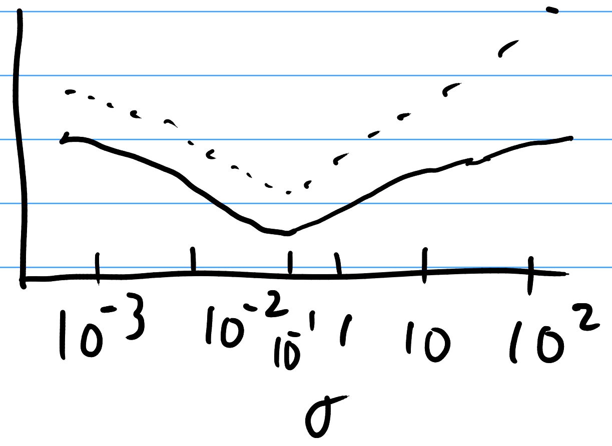
Office hr tomorrow 11 - noon

Homework Q's ?

1.4 $$\nabla_w L = \frac{-1}{n}\sum_{i=1}^{n} \underbrace{P[Y_i = -y_i \mid \vec{x}_i]}_{\text{need practical form}} \vec{x}_i y_i$$

2/3 $$\min_{\vec{w},b} \underbrace{L_{logistic}(\vec{w}, b) + \lambda \|\vec{w}\|^2}_{\text{cost = objective function}}$$

3



after convergence

$$\begin{pmatrix} \vec{w} \\ b \end{pmatrix}$$

loop

evaluate gradient at $\vec{w}_t, b_t$

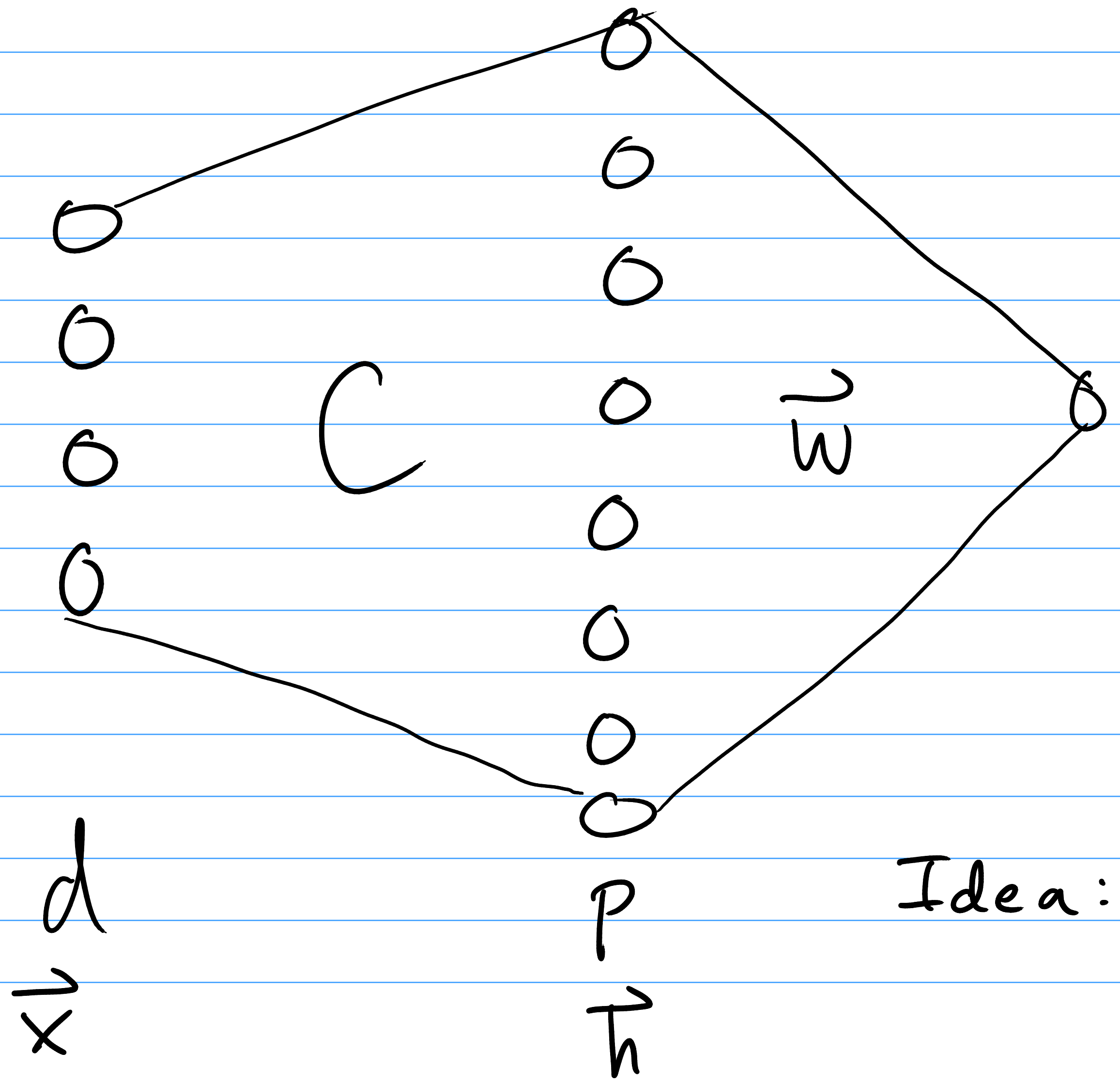
$\vec{w}_{t+1} \leftarrow \vec{w}_t + h \cdot \text{grad } w$

$b_{t+1} \leftarrow b_t + h \cdot \text{grad } b$

$\| \vec{w}_{t+1} - \vec{w}_t \|_\infty$

Random Features

Rahimi
Recht 2008
Neal, Williams
1990's



2-layer NN

$$\vec{h} = g(C\vec{x})$$

$$h_i = g(\vec{c}_i^{\,T}\vec{x})$$

↑ i-th row

$$f(\vec{x}) = \sum_{i=1}^{p} w_i h_i$$

Idea: Think about $p \to \infty$
wide limit

Fix C to its random init

Going to study geometry of hidden layer representations

inner products

Take $\vec{x}, \vec{x}\,' \in \mathbb{R}^d$ two inputs

Study $\frac{1}{p} \underbrace{\vec{h}(\vec{x})}_{\in \mathbb{R}^p}{}^T \vec{h}(\vec{x}\,') = \frac{1}{p} \sum_{i=1}^{p} h_i(\vec{x})\, h_i(\vec{x}\,')$

$$= \frac{1}{p} \sum_{i=1}^{p} g(\vec{c}_i{}^T \vec{x})\, g(\vec{c}_i{}^T \vec{x}\,')$$

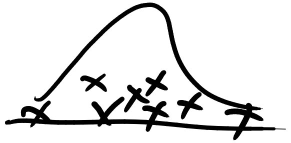
random

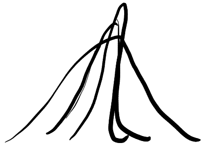
Want to
take $p \to \infty$ ... renormalize sum

think of $h_i \longrightarrow \frac{1}{\sqrt{p}} h_i$

like rescaling c's to be $\mathcal{O}(\frac{1}{\sqrt{p}})$

Law of Large Numbers (weak law)





Collection of random vars X_i : i.i.d. $i = 1, \ldots, p$

$\in \mathbb{R}$

$Var[X_i] = \sigma^2 < \infty$

$\mathbb{E}[(X_i - \mathbb{E}[X_i])^2]$

Partial sum

random var

$$S_p = \frac{1}{p} \sum_{i=1}^{p} X_i \xrightarrow[p \to \infty]{\text{in probability}} \mathbb{E}[X_i]$$

not random

$$\hat{S}_p = \frac{1}{p} \sum_{i=1}^{p} s_i \qquad \mathbb{E}[X_i] = \int X_i \, dP(X_i)$$

$$Pr[\,|S_p - \mathbb{E}[X_i]| > \varepsilon\,] \searrow 0 \quad \text{as } p \to \infty$$

for all $\varepsilon > 0$

Concentration inequality (ML, probability useful)

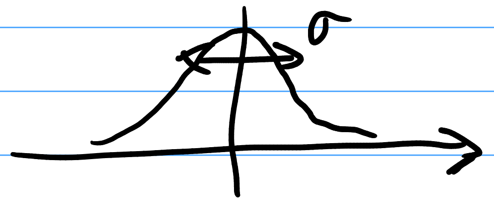
Markov, Hoeffding, etc.

<u>Chebyshev's inequality</u> X r.v. $Var[X] < \infty$

$$\Pr\left[\,|X - \mathbb{E}[X]| \geq a\,\right] \leq \frac{Var[X]}{a^2}$$

ex/ if $a = 5 \cdot \sqrt{Var[X]}$

r.h.s. $= \dfrac{Var[X]}{\left(5 \cdot \sqrt{Var[X]}\right)^2} = \dfrac{1}{25}$



Use Chebyshev on $S_p = \frac{1}{p}\sum_{i=1}^{p} X_i$

Compute $\mathbb{E}[S_p] = \mathbb{E}\left[\frac{1}{p}\sum_{i=1}^{p} X_i\right] = \frac{1}{p}\sum_{i=1}^{p}\mathbb{E}[X_i] = \mathbb{E}[X_i]$

$$\begin{aligned}
\mathrm{Var}[S_p] &= \mathrm{Var}\left[\sum_{i=1}^{p}\left(\frac{X_i}{p}\right)\right] \\
&= \sum_{i=1}^{p} \mathrm{Var}\left[\left(\frac{X_i}{p}\right)\right] \\
&= \sum_{i=1}^{p} \left(\frac{1}{p}\right)^2 \underbrace{\mathrm{Var}[X_i]}_{\sigma^2} \\
&= p \cdot \left(\frac{1}{p}\right)^2 \cdot \sigma^2 \\
&= \frac{\sigma^2}{p}
\end{aligned}$$

$\mathrm{Var}[A+B] = \mathrm{Var}[A] + \mathrm{Var}[B]$ if A, B indep.

$\mathrm{Var}[zA] = z^2\,\mathrm{Var}[A]$

Variance of sample averages decrease as $\frac{1}{p}$

Using Chebyshev,

$$\Pr\left[\,\left|S_p - \mathbb{E}[X_i]\right| \geq \varepsilon\,\right] \leq \frac{\sigma^2}{p\,\varepsilon^2} \searrow 0$$

Ex/ 99% confidence $\Rightarrow$ $\Pr$[further than ε] $<$ 1%

$$\frac{\sigma^2}{p\,\varepsilon^2} < 0.01 \Rightarrow \varepsilon > \frac{\sigma}{\sqrt{p}}\,\frac{1}{\sqrt{0.01}}$$

tolerance $\sim \dfrac{1}{\sqrt{p}}$

Returning to random features

iid $\vec{c}_i$

$$\frac{1}{p}\vec{h}(\vec{x})^T\vec{h}(\vec{x}') = \frac{1}{p}\sum_{i=1}^{p} g(\vec{c}_i^T\vec{x})\, g(\vec{c}_i^T\vec{x}')$$

$$\xrightarrow{P} \mathbb{E}_{\vec{c}}\left[g(\vec{c}^T\vec{x})\, g(\vec{c}^T\vec{x}')\right]$$

$$:= K(\vec{x},\vec{x}')$$

random feature kernel

geometry of network

kernel like in ridge regression / SVM

works for g not too crazy (smooth)

$\vec{c}_i \overset{iid}{\sim} \mu$

$\mathbb{E}\left[\|\vec{c}\|^2\right] < \infty$