Winnowing Subspaces

Manfred Warmuth

University of California - Santa Cruz

ICML 2007, June 21 - 2007

Plots by Dima Kuzmin

To Winnow

How I do it

A formal definition

To winnow:

- to remove (chaff from grain)
  by a current of air
- to get rid of (something undesirable or unwanted)
- From Anglo-Saxon “windwian”

The original Winnow algorithm [L]

- Online alg. for learning disjunctions
- Mistake bound logarithmic in number of features
  winnows a large number of features
- Will morph into algorithm for winnowing subspaces

Disjunctions as linear threshold functions

- 2 out of 5 literal monotone disjunction $v_1 \vee v_3$
- Represented as $\mathbf{d} = (1, 0, 1, 0, 0)^\top$
- Label for instance $\mathbf{x} = (0, 1, 1, 0, 0)^\top$

$$\begin{cases} +1 & \text{if } \mathbf{d} \cdot \mathbf{x} \geq \frac{1}{2} \\ -1 & \text{otherwise} \end{cases}$$

- Alg. receives sequence of examples online

$$\underset{\hat{y}_1}{(\mathbf{x}_1, y_1)} \quad \underset{\hat{y}_2}{(\mathbf{x}_2, y_2)}, \quad \ldots, \quad \underset{\hat{y}_T}{(\mathbf{x}_T, y_T)}$$

instances $[0, 1]^n$, labels and predictions are ± 1

Original Winnow algorithm

Initialize $\mathbf{w}_1 = w_0 \, (1, 1, \ldots 1)^\top$

for $t = 1$ to T **do**

Receive instance $\mathbf{x}_t \in [0, 1]^n$

Predict with

$$\hat{y}_t = \begin{cases} +1 & \text{if } \mathbf{w}_t \cdot \mathbf{x}_t \geq \theta \\ -1 & \text{otherwise} \end{cases}$$

Receive label $y_t \in \{+1, -1\}$

Update

$$w_{t+1,i} = \begin{cases} w_{t,i} & \text{if no mistake} \\ w_{t,i} \; e^{\eta \, y_t x_{t,i}} & \text{if mistake} \end{cases}$$

end for

Like perceptron alg., except multiplicative update

Its merits??

- **Thm** If examples consistent with k out of n literal monotone disjunction, then properly tuned Winnow make at most $O(k \ln n)$ mistakes
- Mistake bound logarithmic in dimension n
- Perceptron alg. can make $\Omega(n\,k)$ mistakes

Morphing into the matrix case?

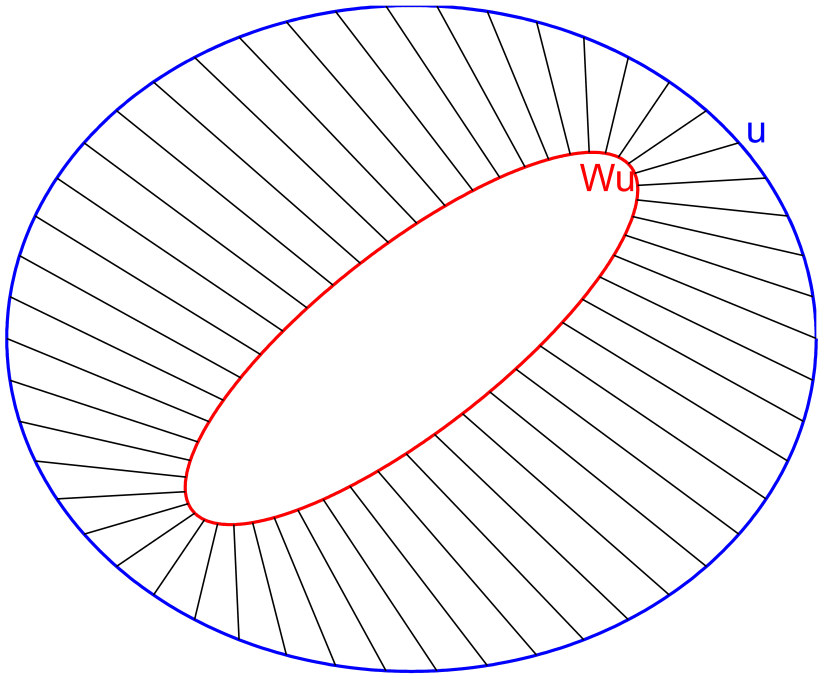
- What variables?
- What dot product?
- What corresponds to disjunctions?
- What happens to the exponential form of weights - soft max?

Lifting it to the matrix setting

	states	mixture states
original	$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	$\mathbf{w} = \begin{pmatrix} 0.1 \\ 0.3 \\ 0.6 \end{pmatrix}$
new	$\mathbf{u}\mathbf{u}^\top$	$\mathbf{W} = \mathbf{U} \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.6 \end{pmatrix} \mathbf{U}^\top$

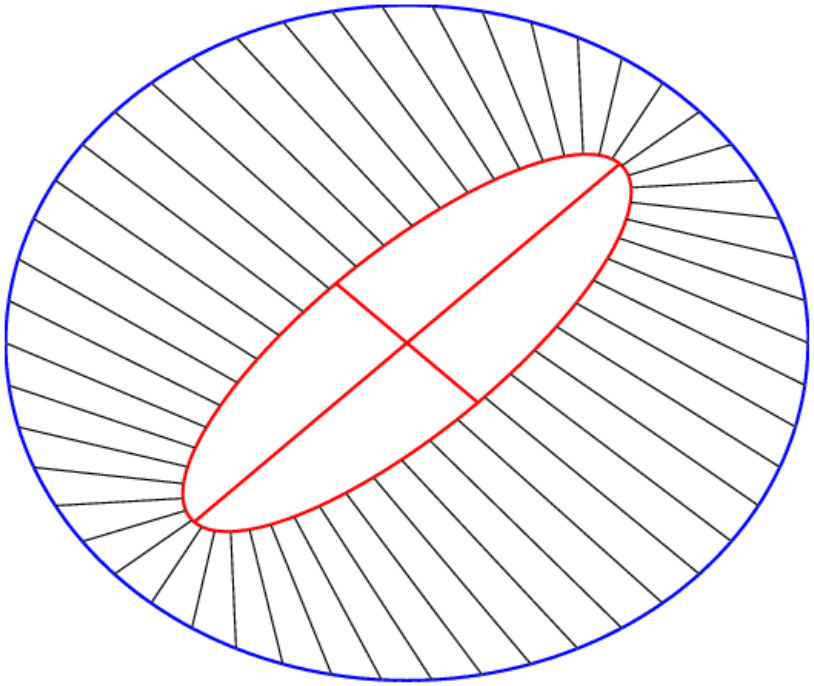
Visualization of symmetric matrices

- As ellipses - affine transformations of the unit ball



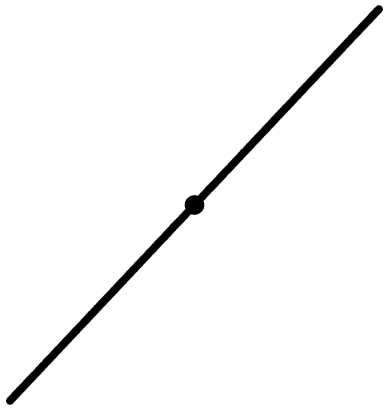
- Ellipse $= \{\mathbf{Wu} : \|\mathbf{u}\|_2 = 1\}$

Ellipses cont.



- Eigenvectors form the axes
  and eigenvalues their lengths

Dyads



Degenerate ellipses

- One eigenvalue one
- All others zero

Linear combinations and mixtures of dyads

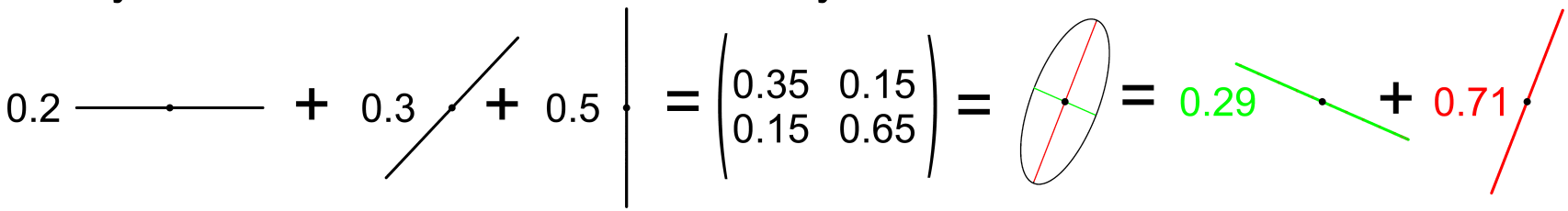
- Symmetric matrices are linear combinations of dyads

$$\underbrace{\mathbf{U}}_{\substack{\text{orthogonal mat.}\\ \text{of eigenvectors}}} \underbrace{\boldsymbol{\lambda}}_{\substack{\text{diagonal mat. of}\\ \text{real eigenvalues}}} \mathbf{U}^\top = \sum_i \underbrace{\lambda_i}_{\substack{\text{real}\\ \text{eigenvalues}}} \underbrace{\mathbf{u}_i \mathbf{u}_i^\top}_{\substack{\text{dyads}}}$$

- Positive definite matrices
  Eigenvalues are non-negative
- Density matrices are mixtures of dyads
  Eigenvalues form probability vector

Mixtures of dyads

- Many mixtures lead to same density matrix



- There always exists a decomposition into n dyads that correspond to eigenvectors
- Uncertainty about dyad expressed as density matrix
- We have a Bayes rule for density matrices

Variance

- View the symmetric positive definite matrix $\mathbf{W}$ as a covariance matrix of some random cost vector $\mathbf{c} \in \mathbf{R}^n$

$$\mathbf{W} = \mathbb{E}\left((\mathbf{c} - \mathbb{E}(\mathbf{c}))(\mathbf{c} - \mathbb{E}(\mathbf{c}))^\top\right)$$

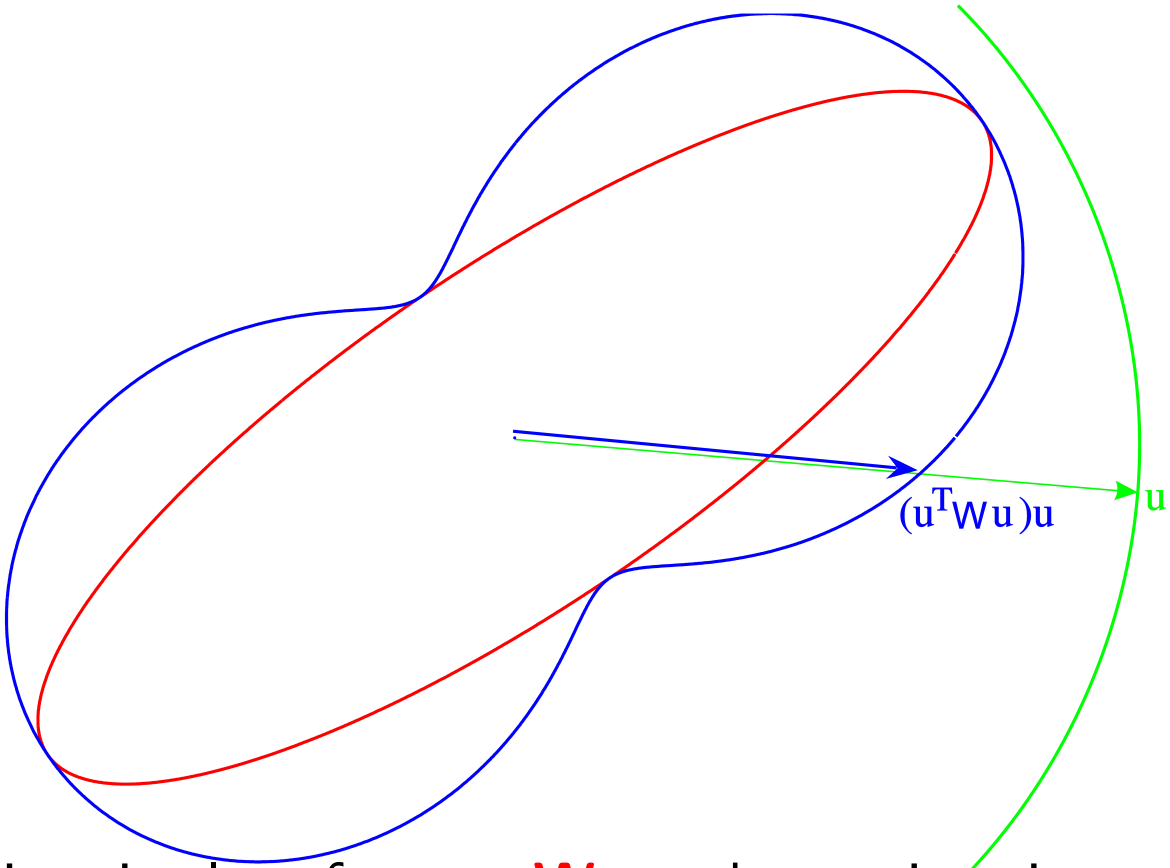
- The variance along any vector $\mathbf{u}$ is

$$\begin{aligned}\mathbf{V}(\mathbf{c}^\top\mathbf{u}) &= \mathbb{E}(\left(\mathbf{c}^\top\mathbf{u} - \mathbb{E}(\mathbf{c}^\top\mathbf{u})\right)^2) \\ &= \mathbf{u}^\top \underbrace{\mathbb{E}\left((\mathbf{c} - \mathbb{E}(\mathbf{c}))(\mathbf{c} - \mathbb{E}(\mathbf{c}))^\top\right)}_{\mathbf{W}} \mathbf{u}\end{aligned}$$

- Variance as trace

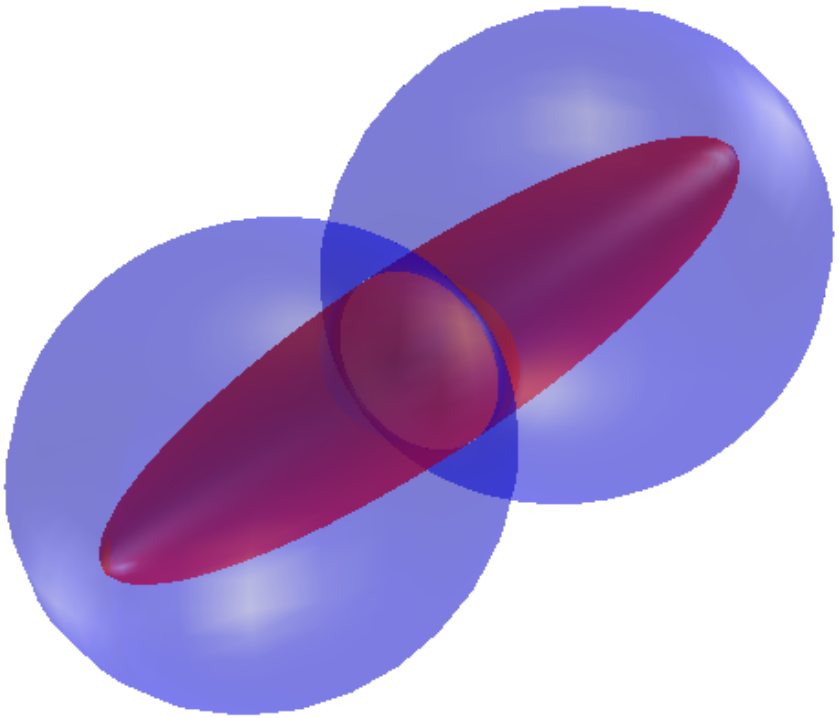
$$\mathbf{u}^\top\mathbf{W}\mathbf{u} = \mathrm{tr}(\mathbf{u}^\top\mathbf{W}\mathbf{u}) = \mathrm{tr}(\mathbf{W}\,\mathbf{u}\mathbf{u}^\top) \geq 0$$

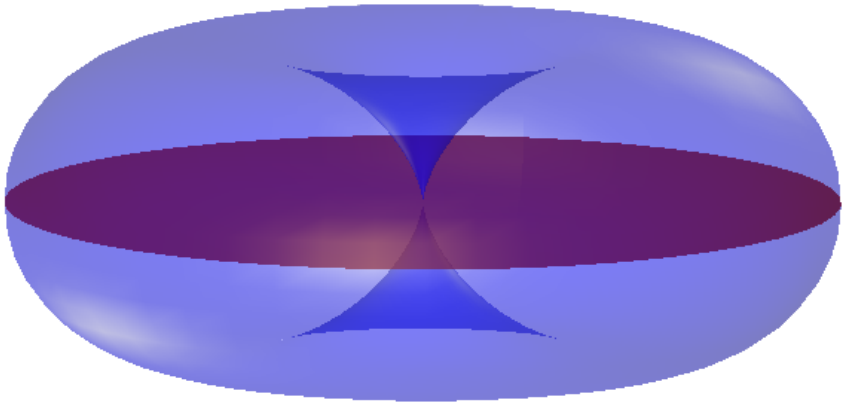
Plotting the variance



Curve of the ellipse is plot of vector **Wu** , where **u** is unit vector
The outer figure eight is direction **u** times the variance $\mathbf{u}^\top \mathbf{W}\mathbf{u}$
For an eigenvector, this variance equals the eigenvalue
and touches the ellipse

3 dimensional variance plots





What dot product?

$$\begin{aligned}
\mathrm{tr}(\mathbf{W}\,\mathbf{X}) &= \mathrm{tr}\Big(\sum_i \omega_i \mathbf{w}_i \mathbf{w}_i^\top\,\mathbf{X}\Big) \\
&= \sum_i \omega_i\,\mathrm{tr}\big(\mathbf{w}_i \mathbf{w}_i^\top\,\mathbf{X}\big) \\
&= \sum_i \omega_i \underbrace{\mathbf{w}_i^\top \mathbf{X} \mathbf{w}_i}_{\text{variance along eigendirs}}
\end{aligned}$$

Measurement in quantum physics

- Dyad $\mathbf{u}\mathbf{u}^\top$ is state
- Density matrix $\mathbf{W}$ is mixture state
- Instance matrix $\mathbf{X}$ is instrument
- $\mathrm{tr}(\mathbf{W}\,\mathbf{X})$ is expected outcome

So far

	states	mixture states	dot product
original	$\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$	$\mathbf{w} = \begin{pmatrix} 0.1 \\ 0.3 \\ 0.6 \end{pmatrix}$	$\mathbf{w} \cdot \mathbf{x}$
new	$\mathbf{u}\mathbf{u}^\top$	$\mathbf{W} = \mathbf{U} \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0.3 & 0 \\ 0 & 0 & 0.6 \end{pmatrix} \mathbf{U}^\top$	$\mathrm{tr}(\mathbf{W}\mathbf{X})$

What corresponds to disjunctions

- Disjunctions

$$(1,0,1,0,0)^\top \cdot (x_1, x_2, x_3, x_4, x_5)^\top = x_1 + x_3$$

Sum k components of $\mathbf{x}$

- Projections matrices $\mathbf{P} = \sum_{i=1}^{k} \mathbf{u}_i \mathbf{u}_i^\top$

$$\mathrm{tr}(\mathbf{PX}) = \sum_{i=1}^{k} \mathbf{u}_i^\top \mathbf{X} \mathbf{u}_i$$

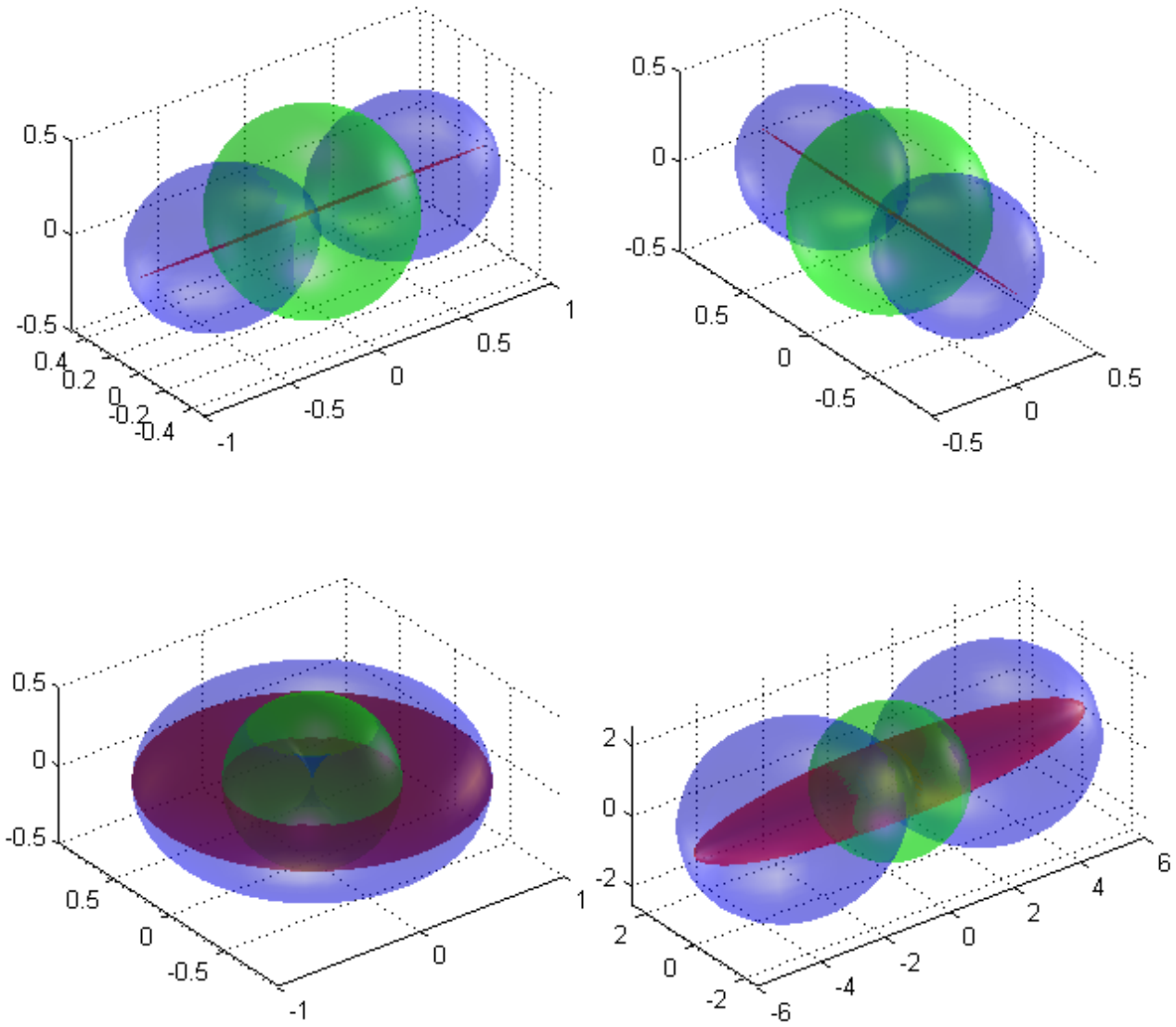
Sum variance along k directions

-

$$(\mathbf{X}_1, \underset{\hat{y}_1}{y_1}), \quad (\mathbf{X}_2, \underset{\hat{y}_2}{y_2}), \quad \ldots, \quad (\mathbf{X}_T, \underset{\hat{y}_T}{y_T})$$

Label y_t is $+1$ if trace is at least $\frac{1}{2}$ and -1 otherwise

Thresholding the trace for $\mathbf{x}\mathbf{x}^\top$ instances



Symmetric Matrix Winnow

Initialize $\mathbf{W}_1 = w_0 \underset{n \times n}{\mathbf{I}}$

for $t = 1$ to T **do**

Receive instance $\underset{n \times n}{\mathbf{X}_t}$ with eigenvalues in $[0, 1]$

Predict with

$$\hat{y}_t = \begin{cases} +1 & \text{if } \mathrm{tr}(\mathbf{W}_t \mathbf{X}_t) \geq \theta \\ -1 & \text{otherwise} \end{cases}$$

Receive label y_t

Update

$$\mathbf{W}_{t+1} = \begin{cases} \mathbf{W}_t & \text{if no mistake} \\ \mathbf{exp}(\mathbf{log}\, \mathbf{W}_t + \eta\, y_t \mathbf{X}_t) & \text{if mistake} \end{cases}$$

end for

exp and **log** are spectral functions

In normalized version, trace normalized to one

What else?

- Same $O(k \ln n)$ mistake bound if examples consistent with k-dimensional subspace
  - Dubbed free matrix lunch
- Generalizes to arbitrary matrices
  - use mixtures of $\mathbf{u}\mathbf{v}^\top$ and SVD
- Key tool in analysis is quantum relative entropy [TRW]

$$\Delta(\mathbf{W}, \mathbf{V}) = \mathrm{tr}(\mathbf{W}(\log \mathbf{W} - \log \mathbf{V}))$$