Learning a set of directions

Wouter M. Koolen, Jiazhong Nie and Manfred K. Warmuth

Centrum Wiskunde & Informatica

UNIVERSITY OF CALIFORNIA SANTA CRUZ

COLT

June 12$^{\text{th}}$ 2013

Amsterdam 5 meters below sea level

Pump H_2O - but where to point the Windmills?

Online learning to help: for $t = 1, 2, \ldots$

- Mill chooses a randomized direction $\boldsymbol{u}_t \sim \mathbb{P}_t$
- Wind reveals direction $\boldsymbol{x}_t$
- Expected **gain** based on match

Randomized Prediction

For $\boldsymbol{u} \sim \mathbb{P}$,

$$\mathbb{E}\left[(\boldsymbol{u}^\intercal \boldsymbol{x} + c)^2\right] = \boldsymbol{x}^\intercal \underbrace{\mathbb{E}[\boldsymbol{u}\boldsymbol{u}^\intercal]}_{\text{2nd moment } \boldsymbol{D}} \boldsymbol{x} + 2c\boldsymbol{x}^\intercal \underbrace{\mathbb{E}[\boldsymbol{u}]}_{\text{1st moment } \boldsymbol{\mu}} + c^2$$

Key idea: Use parameter $\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle$

What is set $\mathcal{U}$ of valid $\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle$?

$$\mathcal{U} := \{\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle \mid \exists \mathbb{P} : \boldsymbol{\mu}, \boldsymbol{D} \text{ are 1st/2nd moment of some } \mathbb{P}\}$$

Characterisation Theorem Parameter $\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle \in \mathcal{U}$ iff $\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle$ satisfies the following **semi-definite** constraints:

$$\operatorname{tr}(\boldsymbol{D}) = 1 \quad \text{and} \quad \boldsymbol{D} \succeq \boldsymbol{\mu}\boldsymbol{\mu}^\intercal$$

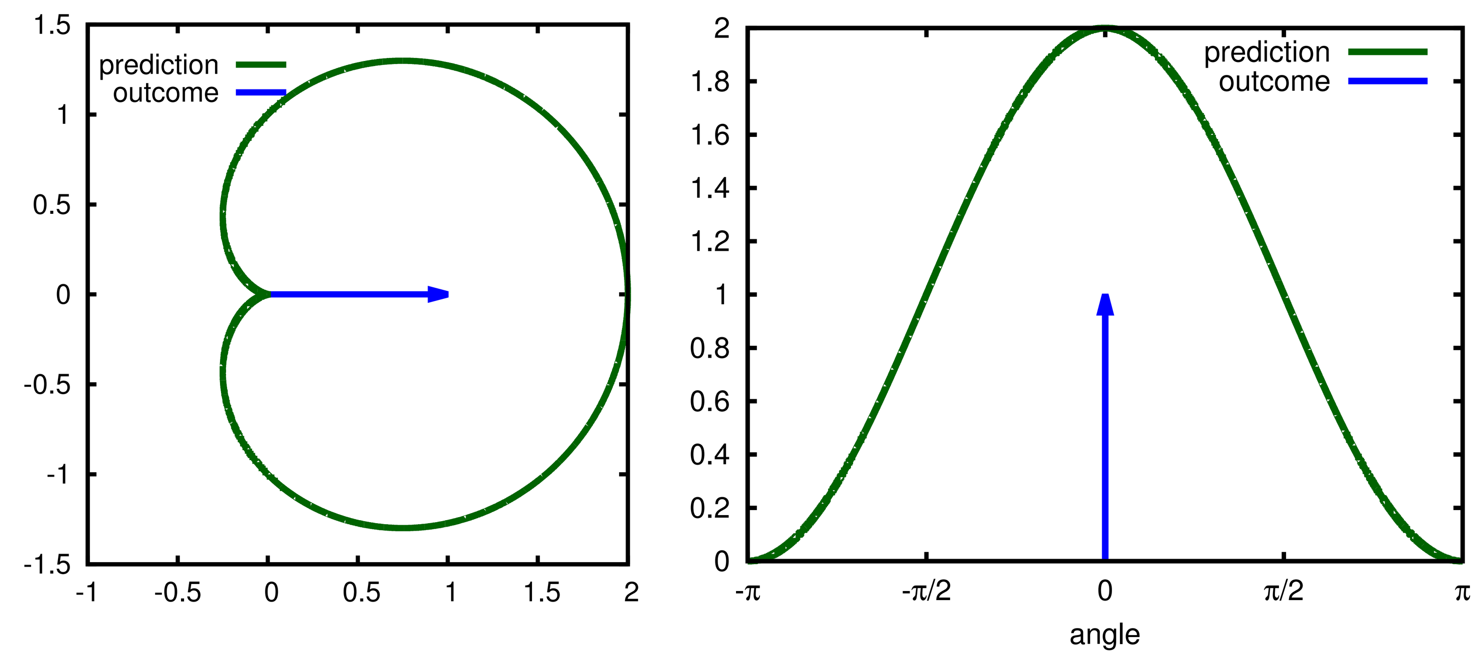
and any $\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle \in \mathcal{U}$ can be efficiently decomposed into $2(n+1)$ "pure" directions:

$$\langle \boldsymbol{\mu}, \boldsymbol{D} \rangle = \sum_{i=1}^{2(n+1)} w_i \langle \boldsymbol{u}_i, \boldsymbol{u}_i \boldsymbol{u}_i^\intercal \rangle$$

What is a reasonable gain?

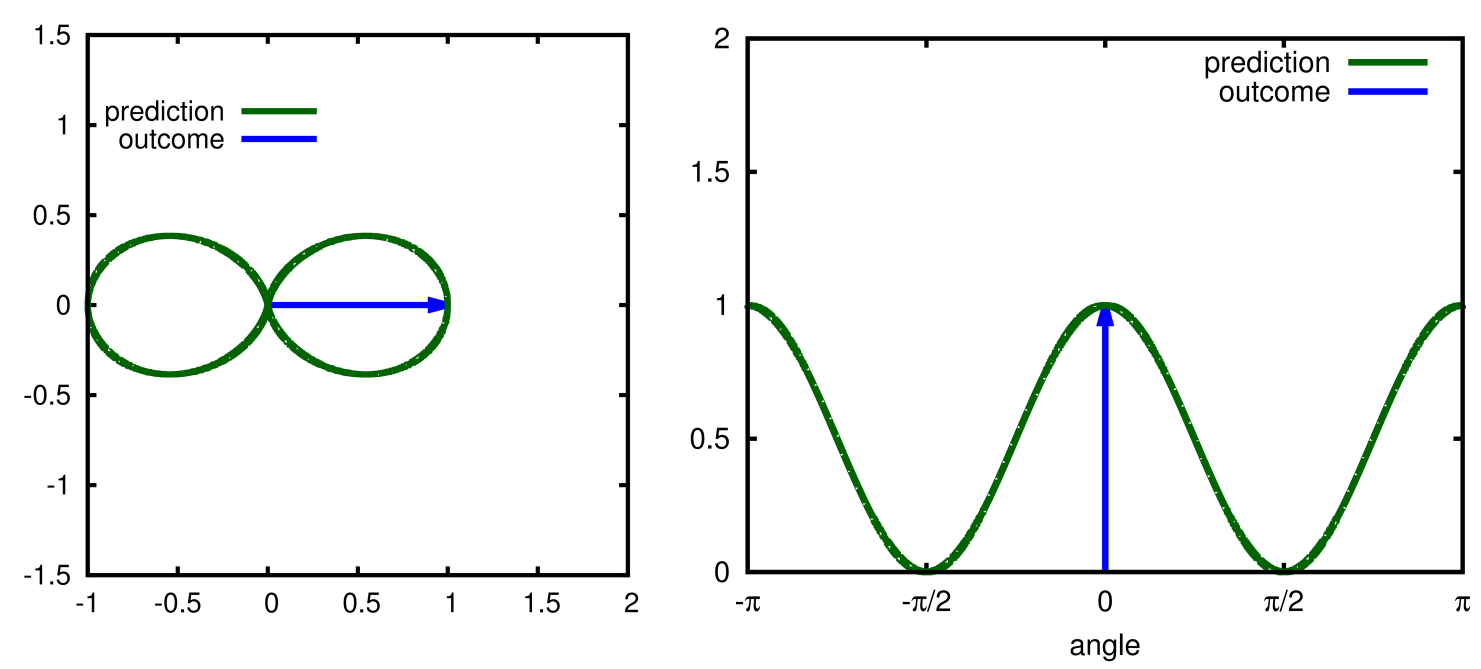
Shifted angle cosine $:= \boldsymbol{u}^\intercal \boldsymbol{x} + 1$

best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel, worst when $\boldsymbol{u}$,$\boldsymbol{x}$ opposite



Subspace similariy gain $:= (\boldsymbol{u}^\intercal \boldsymbol{x})^2$

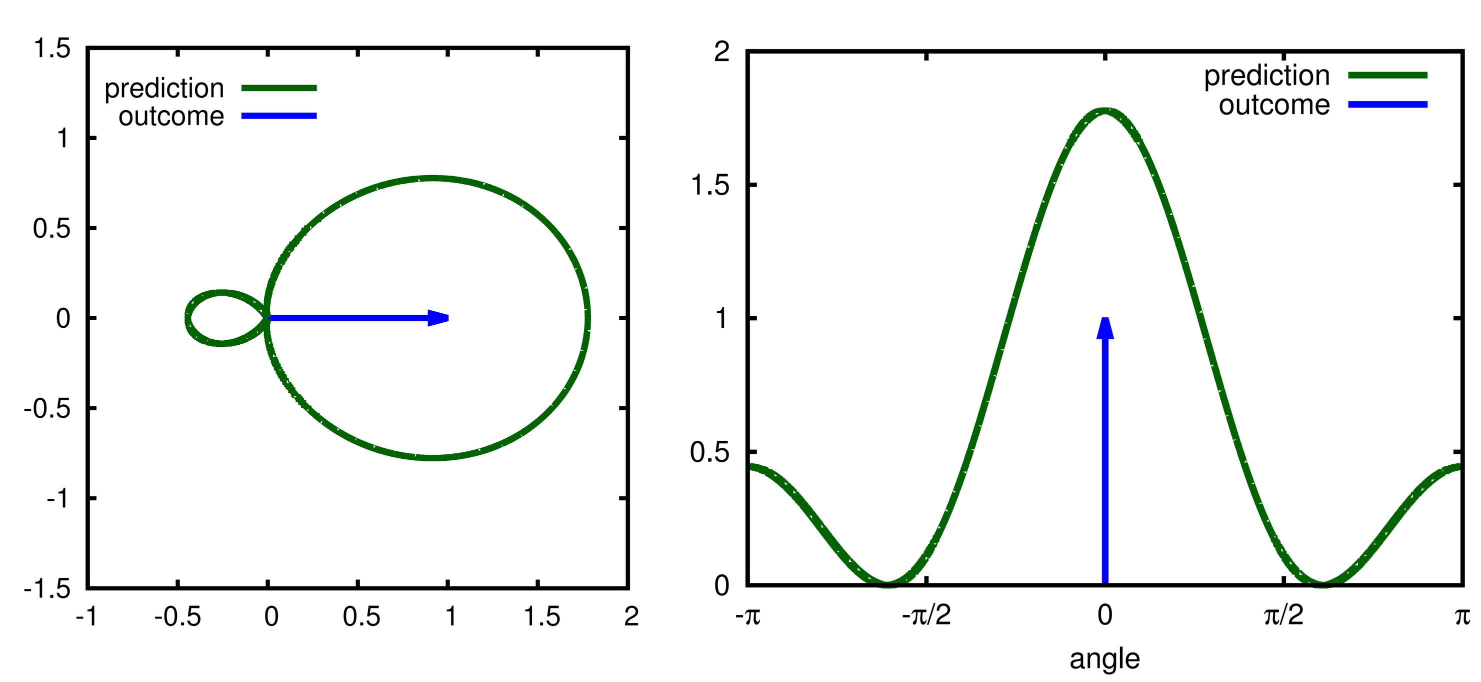
Used in Principle Component Analysis
best when $\boldsymbol{u}$,$\boldsymbol{x}$ parallel or opposite



Our directional gain $:= (\boldsymbol{u}^\intercal \boldsymbol{x} + c)^2$

$$= (\boldsymbol{u}^\intercal \boldsymbol{x})^2 + 2c\boldsymbol{u}^\intercal \boldsymbol{x} + c^2$$

- A tradeoff controlled by mill-dependent constant c
- Quadratic Taylor approximation of any gain at $\boldsymbol{u} = 0$



Tradeoff constant $c = 1/3$

Gradient descent

Mill maintains the two moments $\langle \boldsymbol{\mu}_t, \boldsymbol{D}_t \rangle \in \mathcal{U}$ as parameter
At trial $t = 1 \ldots T$, the Mill

1. **Decomposes** parameter $\langle \boldsymbol{\mu}_t, \boldsymbol{D}_t \rangle$ into a mixture of directions and draws $\boldsymbol{u}_t$ from mixture
2. Receives Wind direction $\boldsymbol{x}_t$ and gain $\mathbb{E}\left[(\boldsymbol{u}_t^\intercal \boldsymbol{x}_t + c)^2\right]$
3. **Updates** $\langle \boldsymbol{\mu}_t, \boldsymbol{D}_t \rangle$ to $\langle \widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{D}}_{t+1} \rangle$ with the gradient of the expected gain on $\boldsymbol{x}_t$
$$\widehat{\boldsymbol{\mu}}_{t+1} := \boldsymbol{\mu}_t + 2\eta c\, \boldsymbol{x}_t \quad \text{and} \quad \widehat{\boldsymbol{D}}_{t+1} := \boldsymbol{D}_t + \eta\, \boldsymbol{x}_t \boldsymbol{x}_t^\intercal$$
4. **Projects** $\langle \widehat{\boldsymbol{\mu}}_{t+1}, \widehat{\boldsymbol{D}}_{t+1} \rangle$ back into $\mathcal{U}$
$$\langle \boldsymbol{\mu}_{t+1}, \boldsymbol{D}_{t+1} \rangle := \operatorname*{argmin}_{\substack{\operatorname{tr}(\boldsymbol{D}) = 1 \\ \boldsymbol{D} \succeq \boldsymbol{\mu}\boldsymbol{\mu}^\intercal}} \|\boldsymbol{D} - \widehat{\boldsymbol{D}}_{t+1}\|_F^2 + \|\boldsymbol{\mu} - \widehat{\boldsymbol{\mu}}_{t+1}\|^2$$

Theorem

With proper tuning of η, the regret after T trials of GD is at most $\sqrt{3(4c^2+1)T}$

- Regret grows sub-linearly with T
- Mill turned close to the best orientation
- Holland is saved ☺

Conclusion

- An efficient method for orienting windmills
- Characterization of set of first two moments of distributions on directions
- Works for $n \geq 2$ dimensions
- We can learn sets of $k \geq 1$ orthogonal directions. Characterisation Theorem and decomposition alg. much more tricky

Bonus plots