Contrastive Moments: Unsupervised Halfspace Learning in Polynomial Time

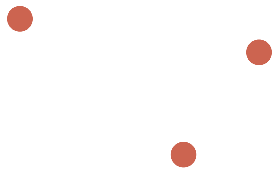
Xinyuan Cao, Santosh S. Vempala

Given data with +/- labels

- positive samples
- negative samples

Given data with +/- labels

- positive samples
- negative samples

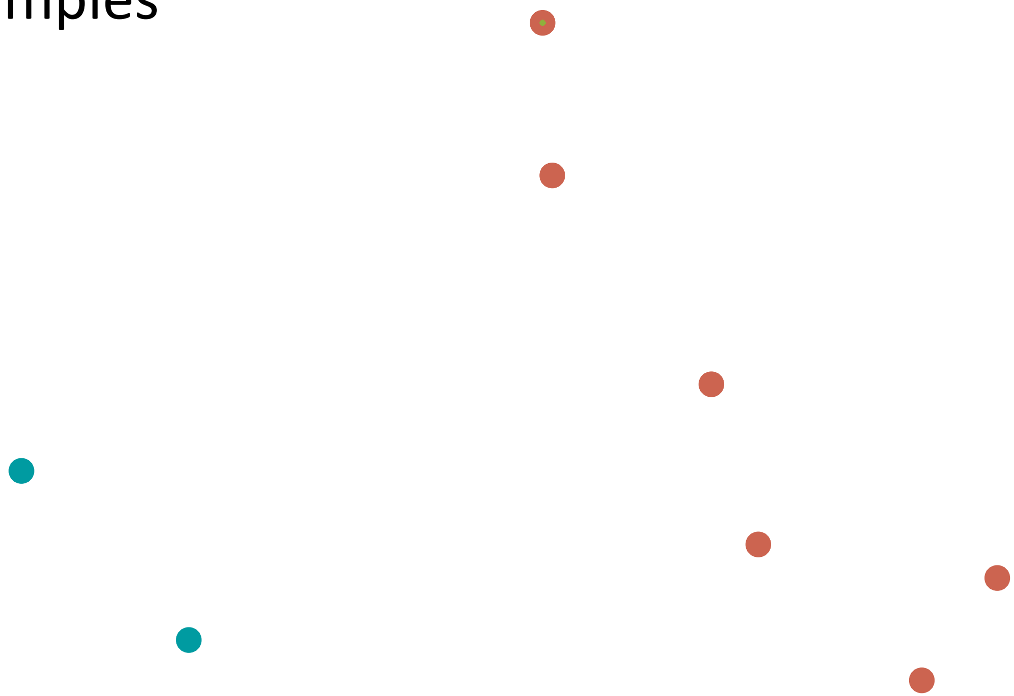


Given data with +/- labels

- positive samples
- negative samples

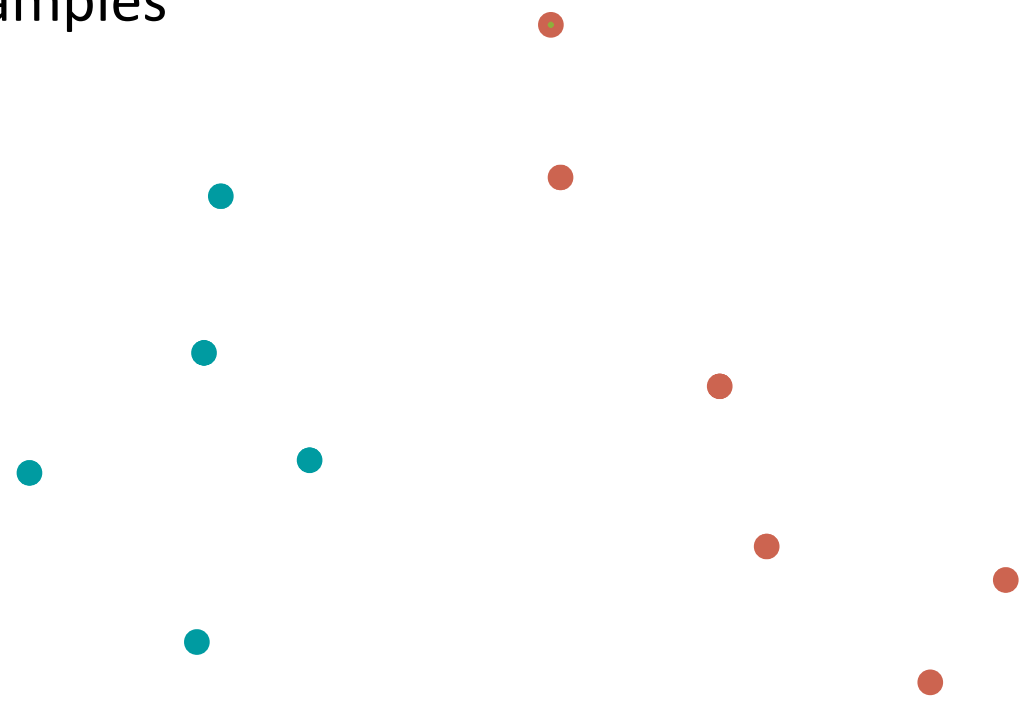
Given data with +/- labels

- positive samples
- negative samples



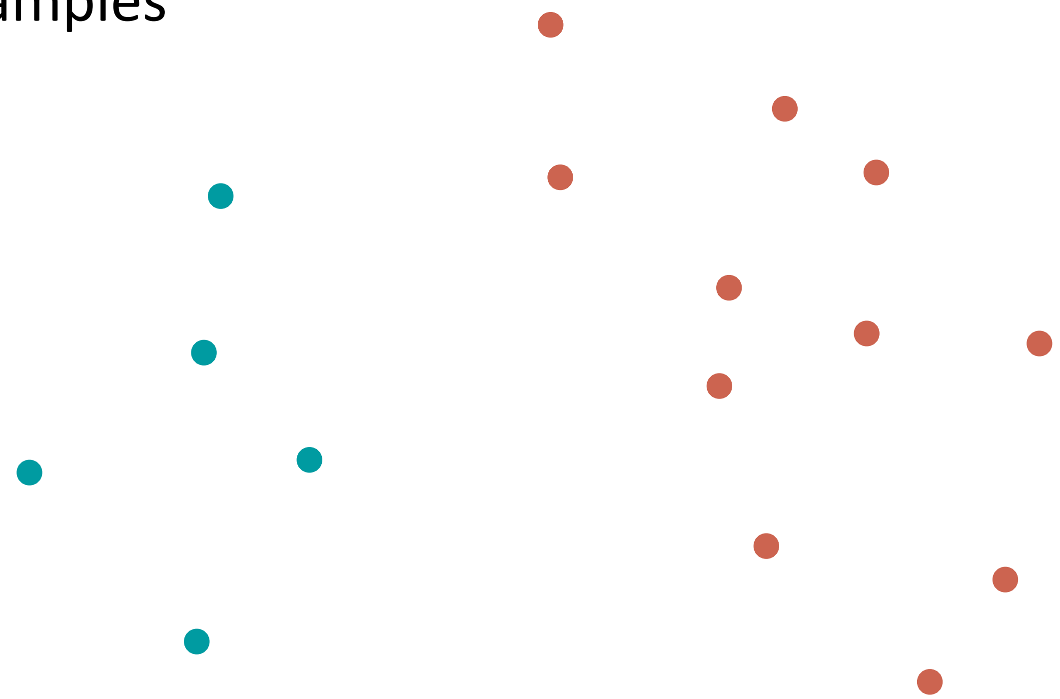
Given data with +/- labels

- positive samples
- negative samples



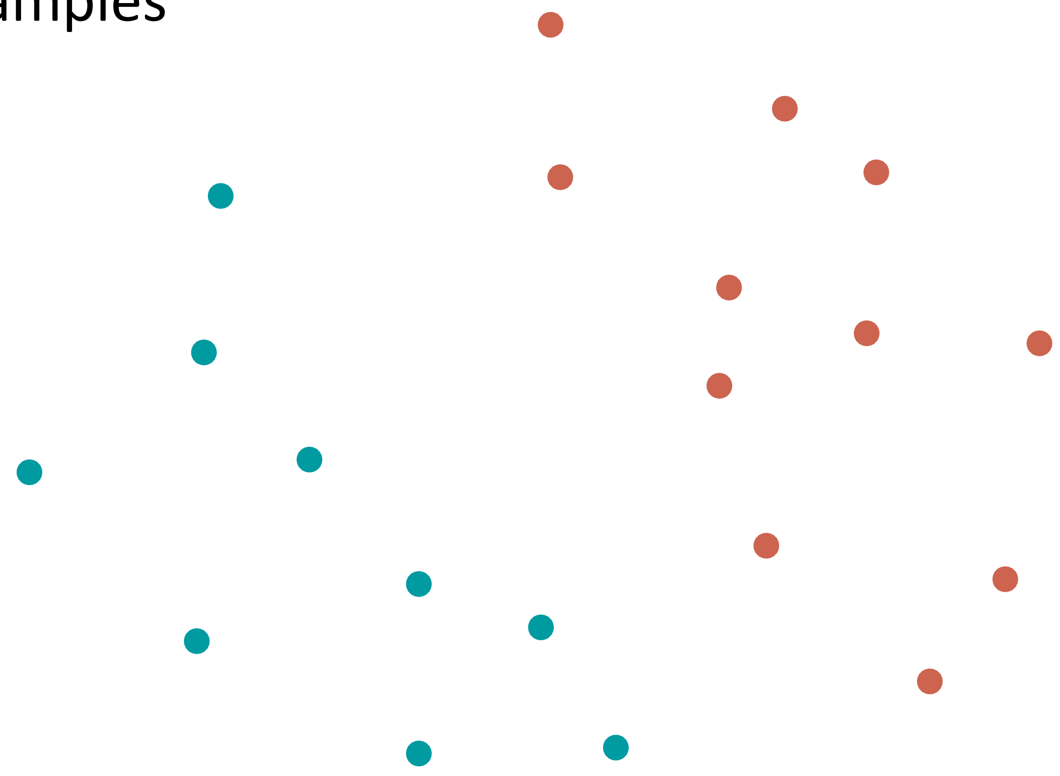
Given data with +/- labels

- positive samples
- negative samples



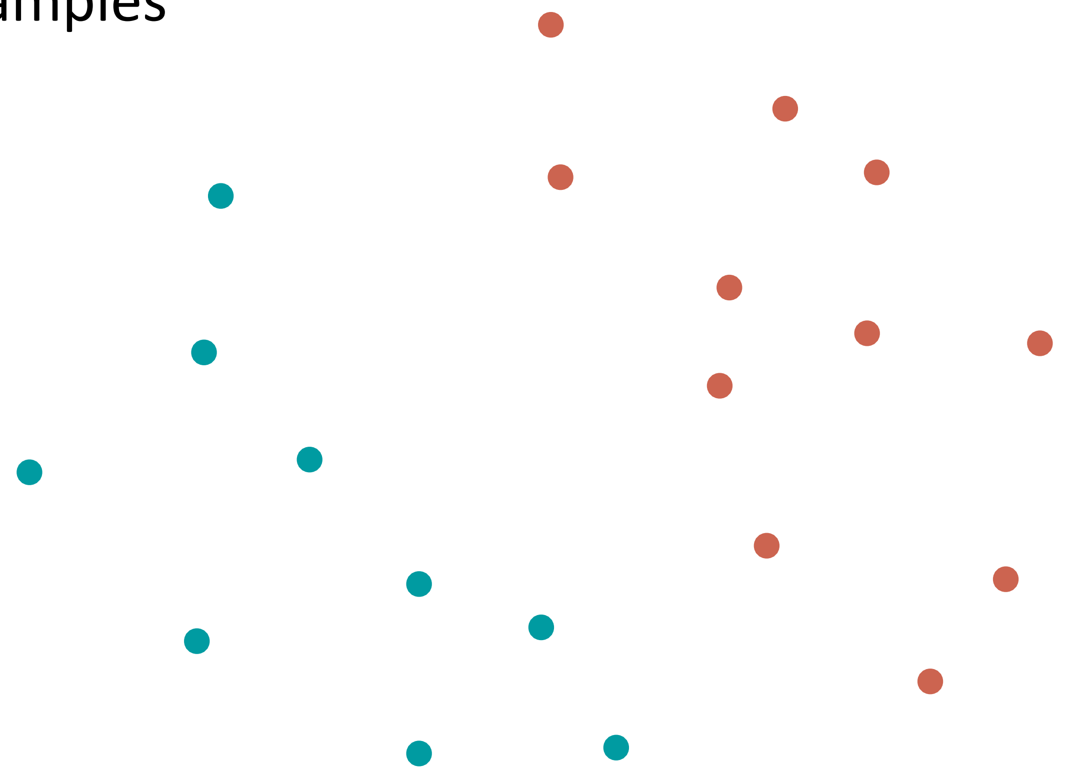
Given data with +/- labels

- positive samples
- negative samples



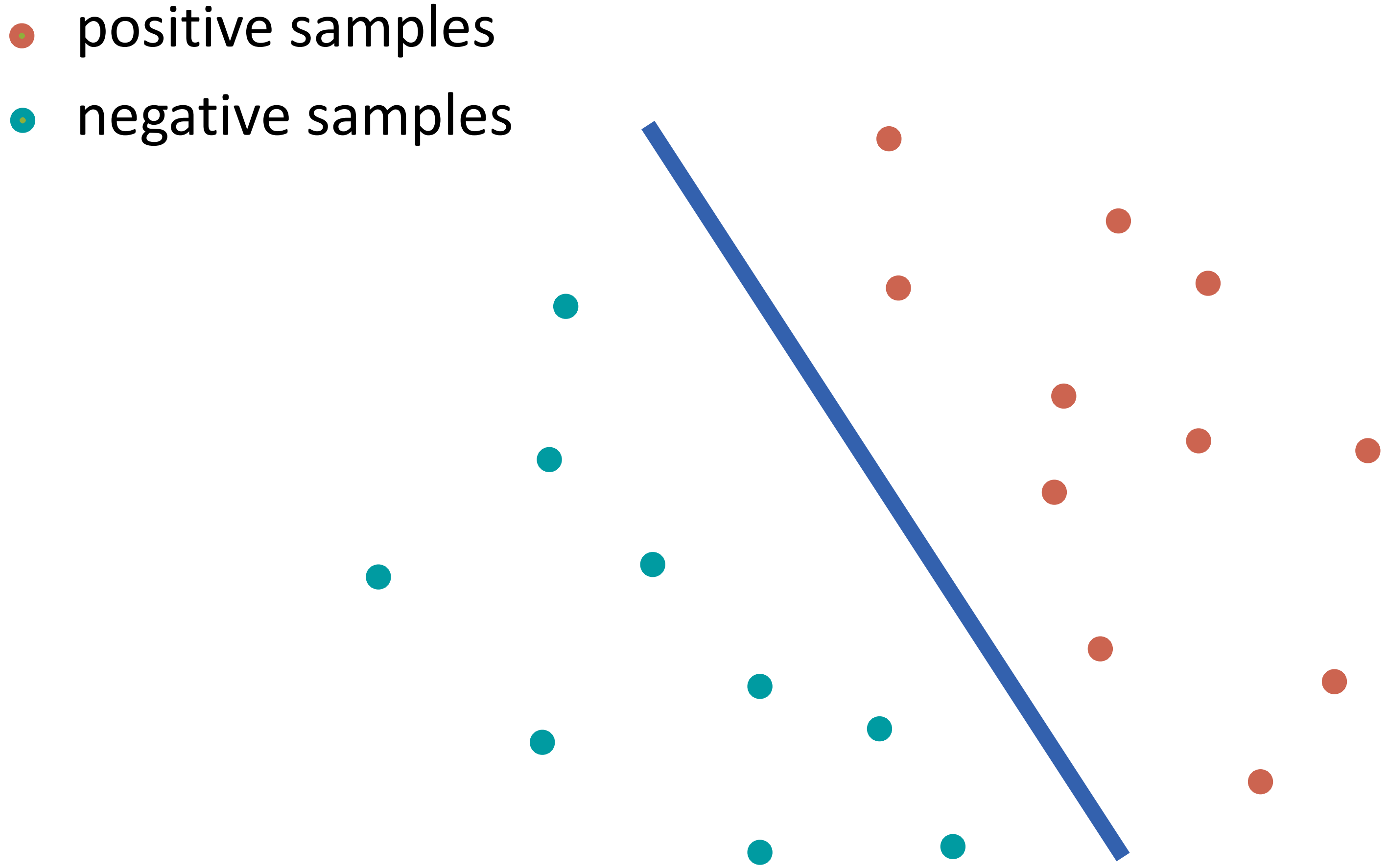
Given data with +/- labels

- positive samples
- negative samples



What can we learn from the labeled data?

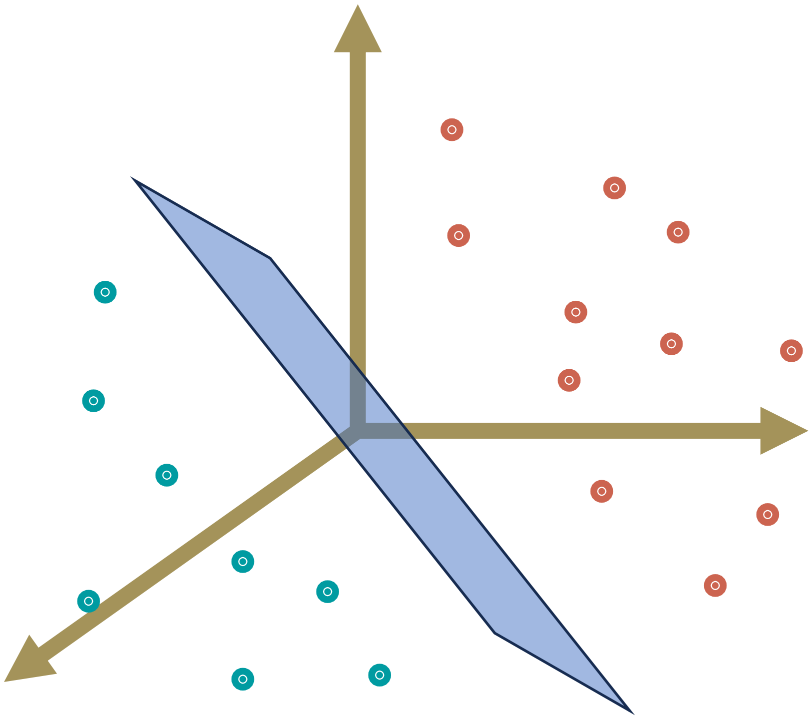
Given data with +/- labels



What can we learn from the labeled data? Halfspace!

Given data with +/- labels

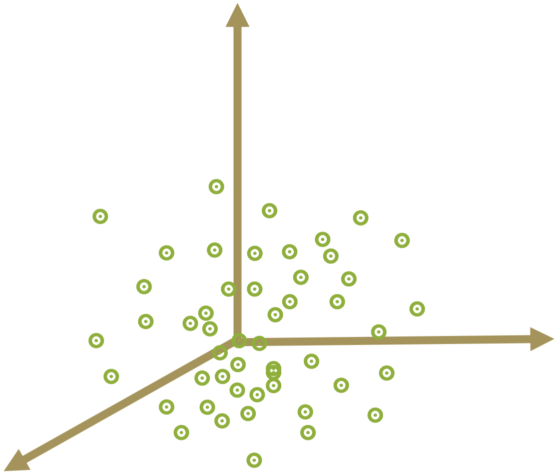
- positive samples
- negative samples



What can we learn from the labeled data? Halfspace!

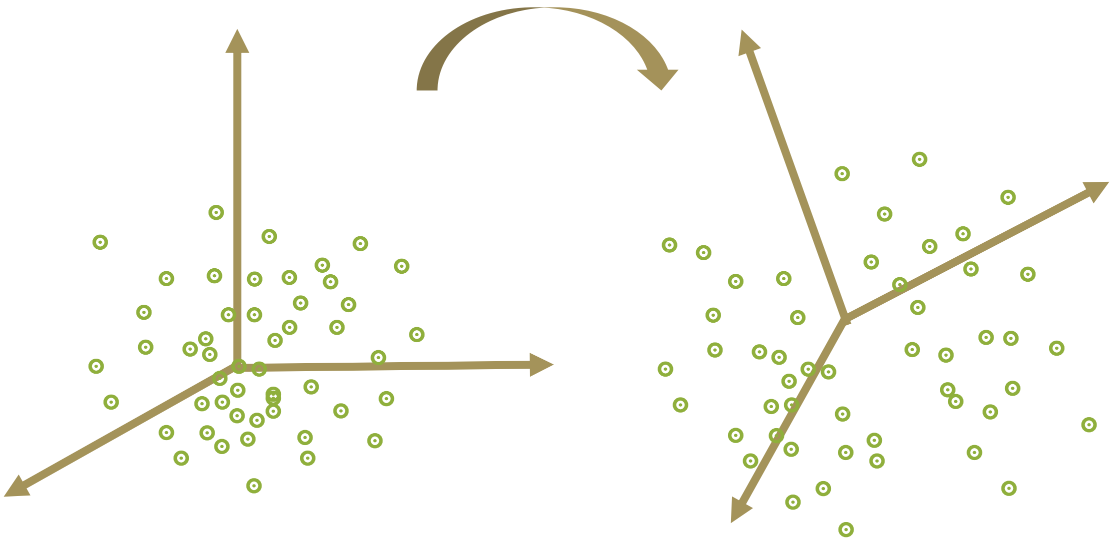
Given data without labels

What can we learn from unlabeled data?



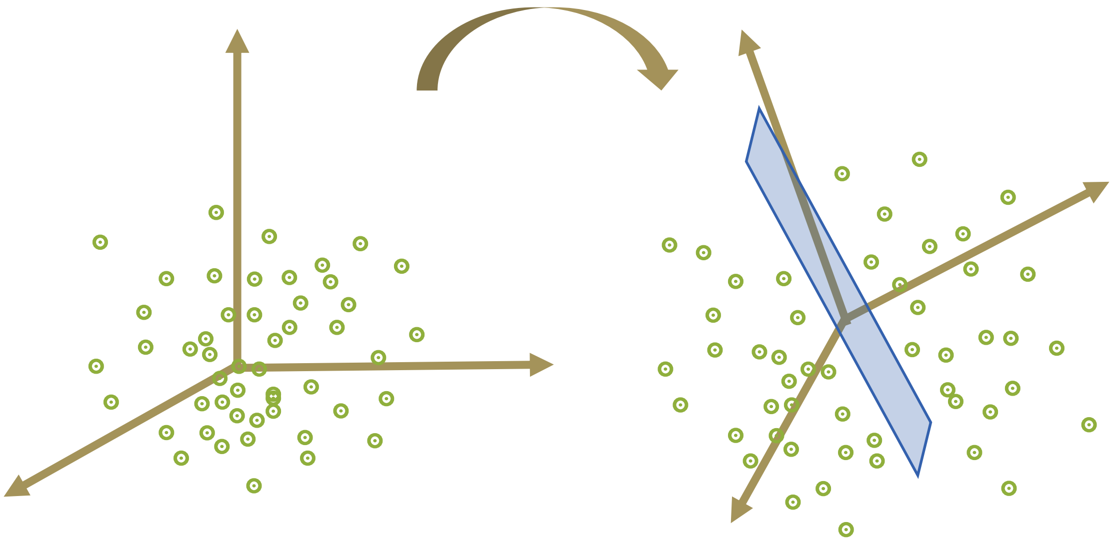
Given data without labels

What can we learn from unlabeled data?



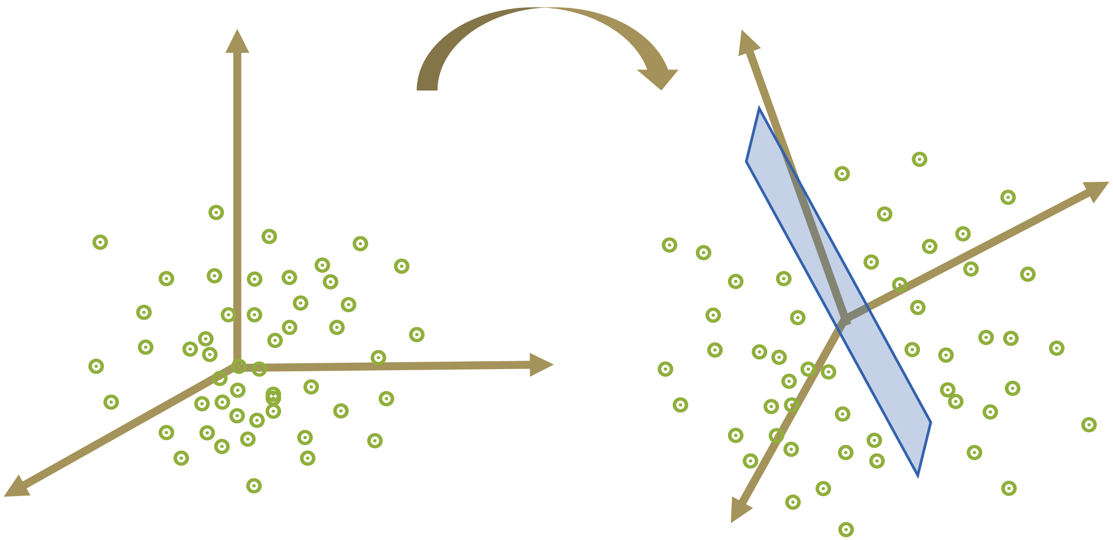
Given data without labels

What can we learn from unlabeled data? Halfspace!



Given data without labels

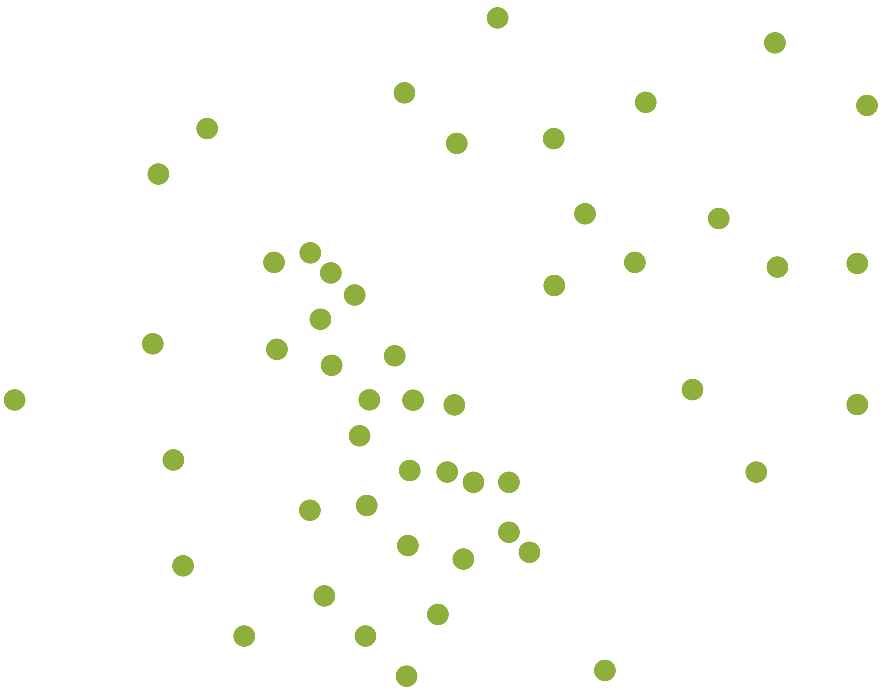
What can we learn from unlabeled data? Halfspace!



Unsupervised Halfspace Learning

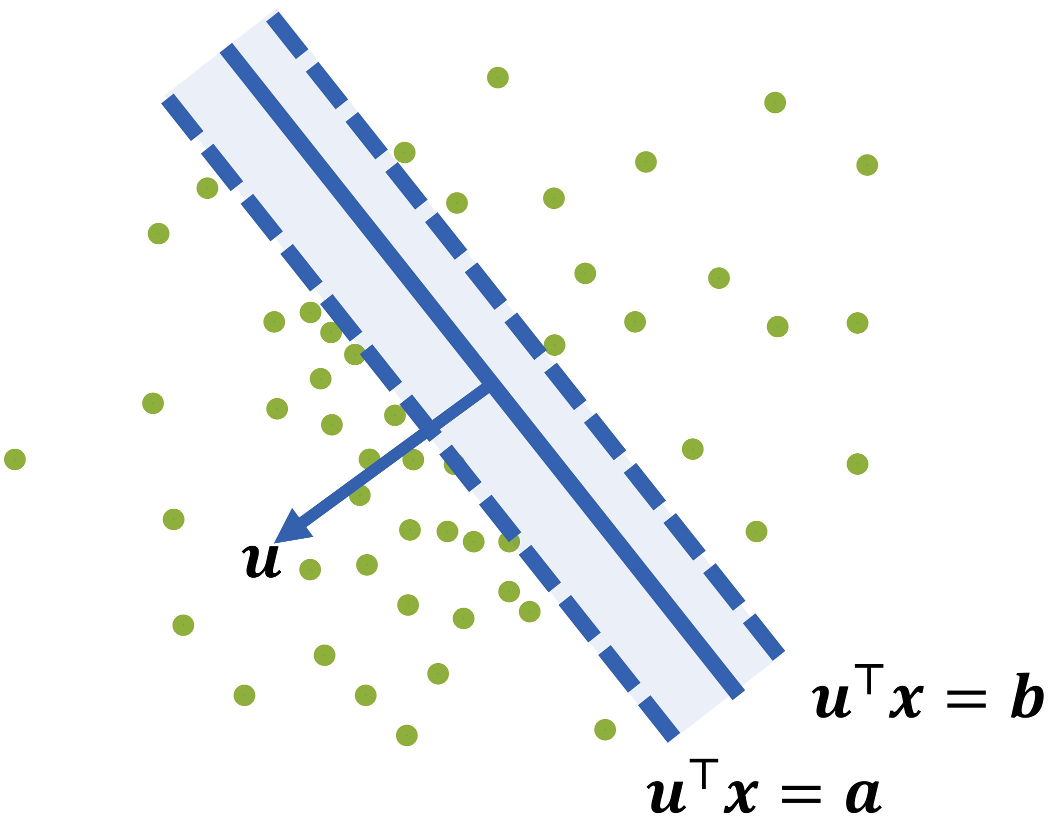
Problem

Input: Unlabeled data from distribution $\widehat{\boldsymbol{P}}$ in $\mathbb{R}^d$.



Problem

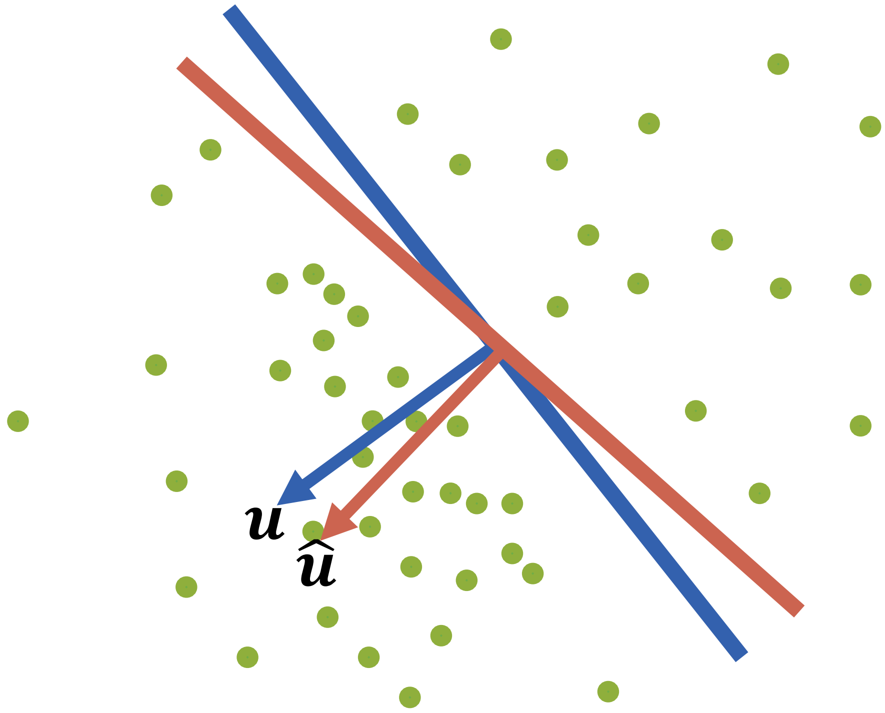
Input: Unlabeled data from distribution $\widehat{\boldsymbol{P}}$ in $\mathbb{R}^d$.
There is an ϵ-margin halfspace.



Problem

Input: Unlabeled data from distribution $\widehat{\boldsymbol{P}}$ in $\mathbb{R}^d$.
There is an ϵ-margin halfspace.

Output: Normal vector $\widehat{\boldsymbol{u}}$ to within TV distance δ.



Problem

Input: Unlabeled data from distribution $\widehat{\boldsymbol{P}}$ in $\mathbb{R}^d$.
There is an ϵ-margin halfspace.

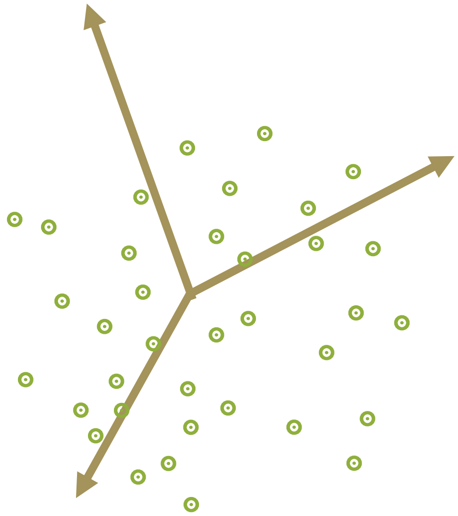
Output: Normal vector $\widehat{\boldsymbol{u}}$ to within TV distance δ.

Main result

There is an algorithm that can learn any **affine product logconcave distribution with ϵ-margin** to within TV distance δ with time and sample complexity that are $\mathbf{poly}(\boldsymbol{d}, \mathbf{1}/\boldsymbol{\varepsilon}, \mathbf{1}/\boldsymbol{\delta})$ whp.

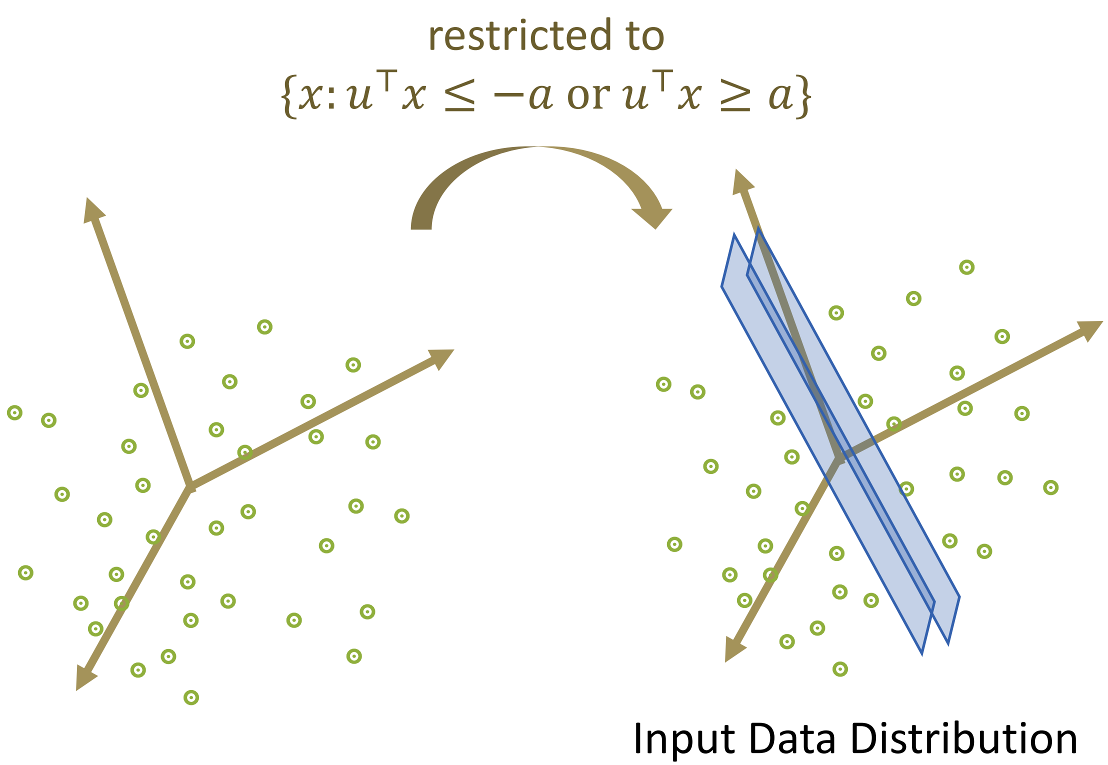
Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity)
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?



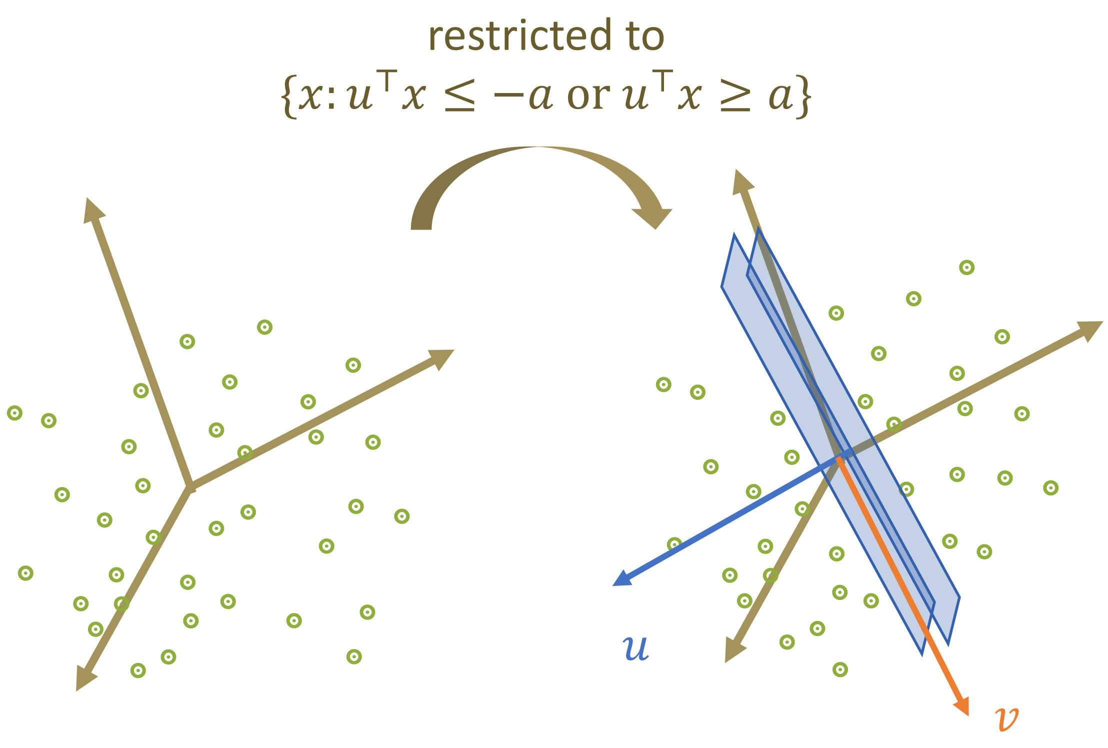
Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?



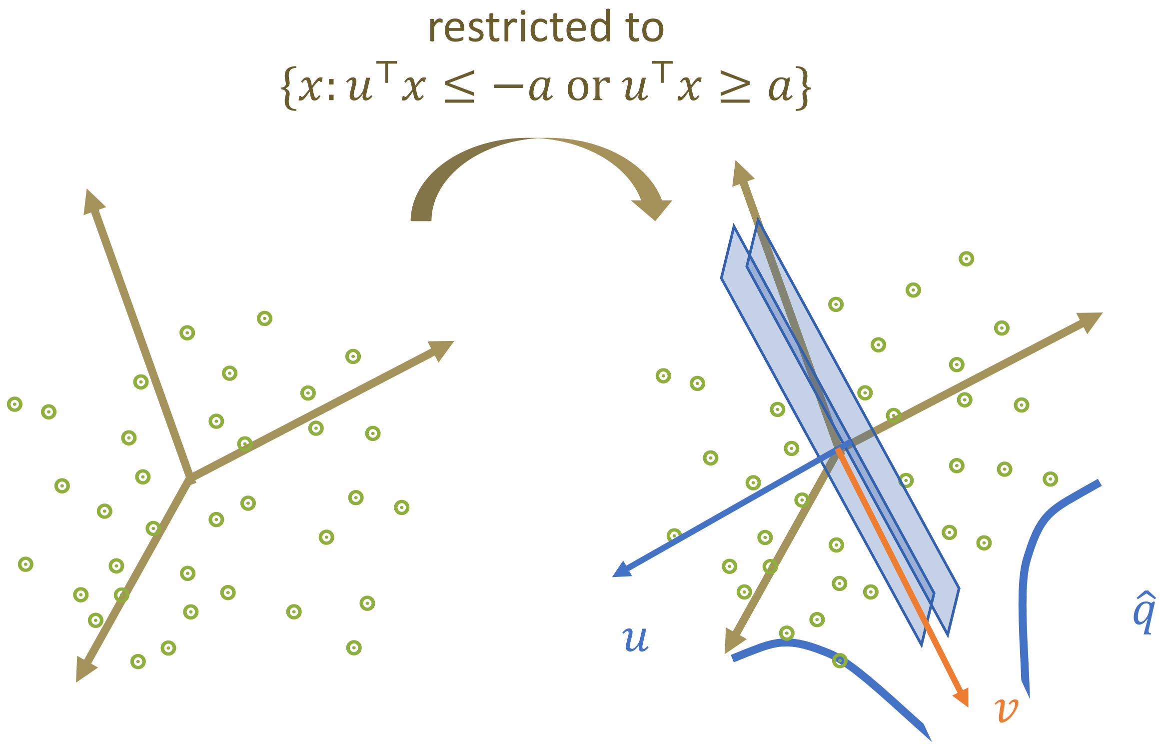
Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?



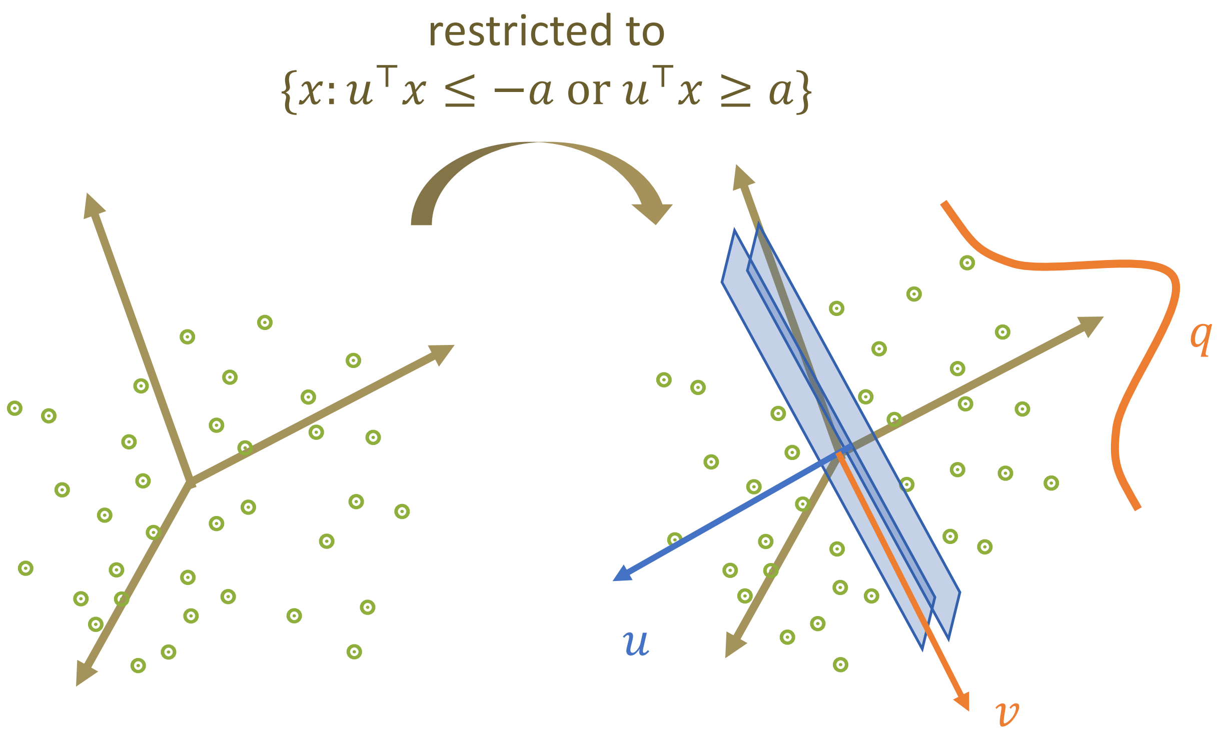
Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?



Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?

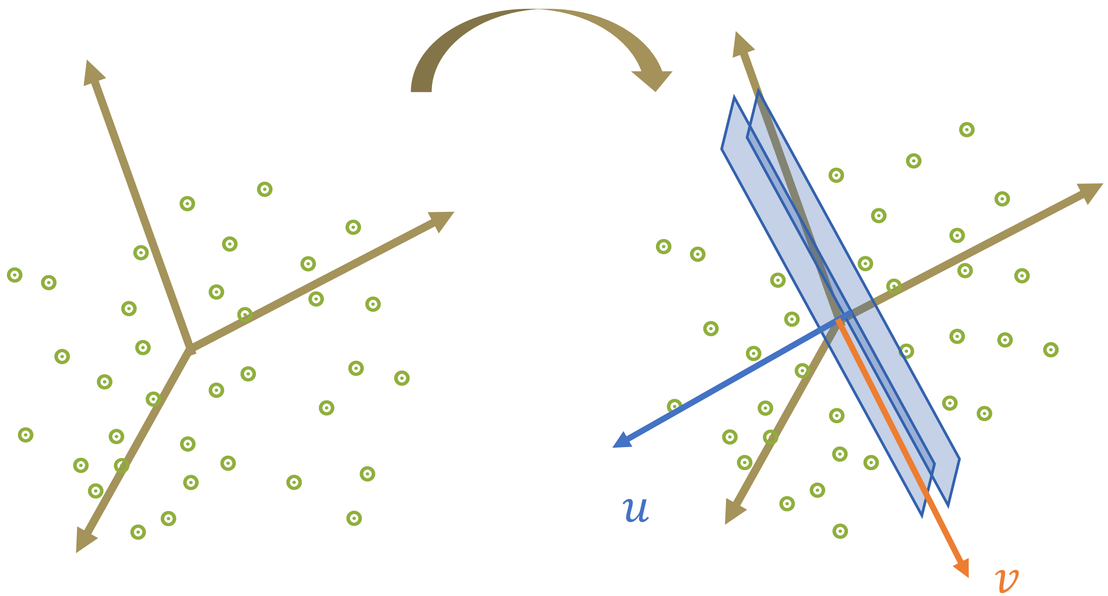


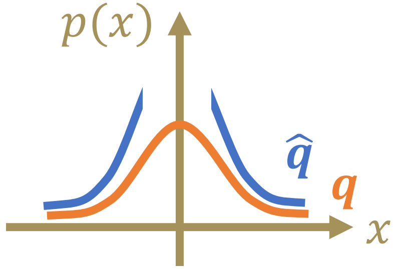
Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?

q: isotropic density
$\hat{q}$: isotropic density with margin

restricted to
$\{x: u^\top x \leq -a \text{ or } u^\top x \geq a\}$



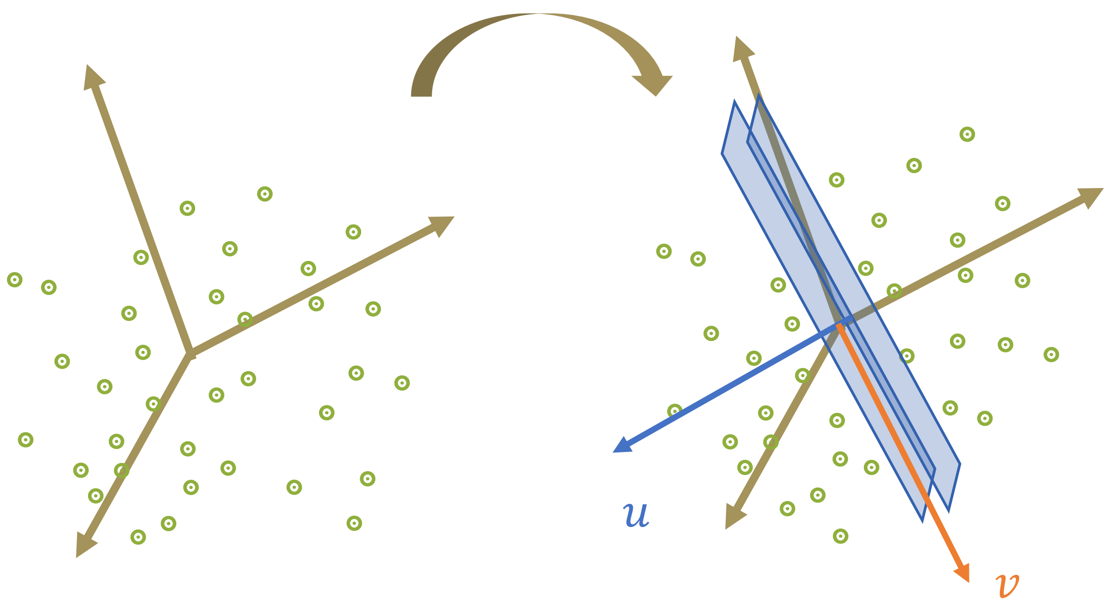


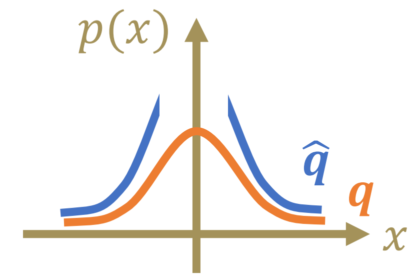
Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?

q: isotropic density
$\hat{q}$: isotropic density with margin

restricted to
$\{x: u^\top x \leq -a \text{ or } u^\top x \geq a\}$

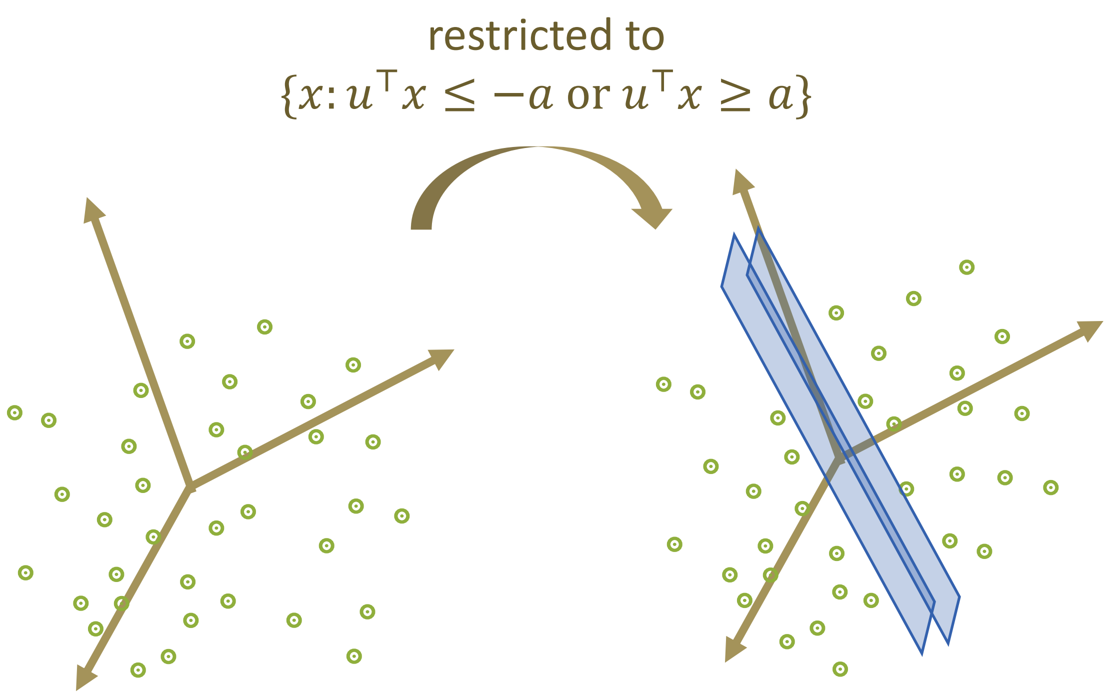




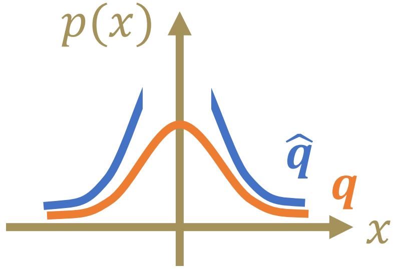
Find the direction that maximizes variance.

Warm-up: isotropic, symmetric margin

- Isotropic distribution (mean zero, covariance identity).
- Margin symmetric with the origin.
- Can we find the halfspace efficiently?



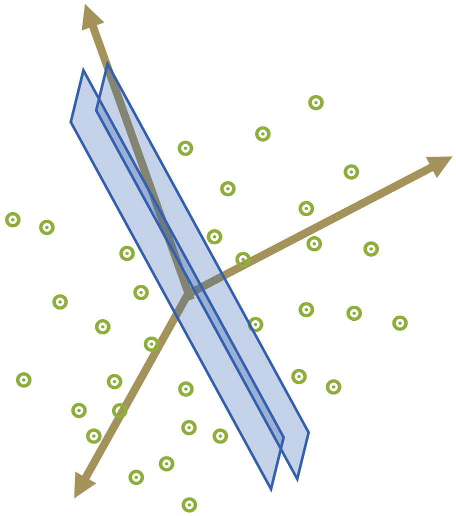
q: isotropic density
$\hat{q}$: isotropic density with margin



Find the direction that maximizes variance.

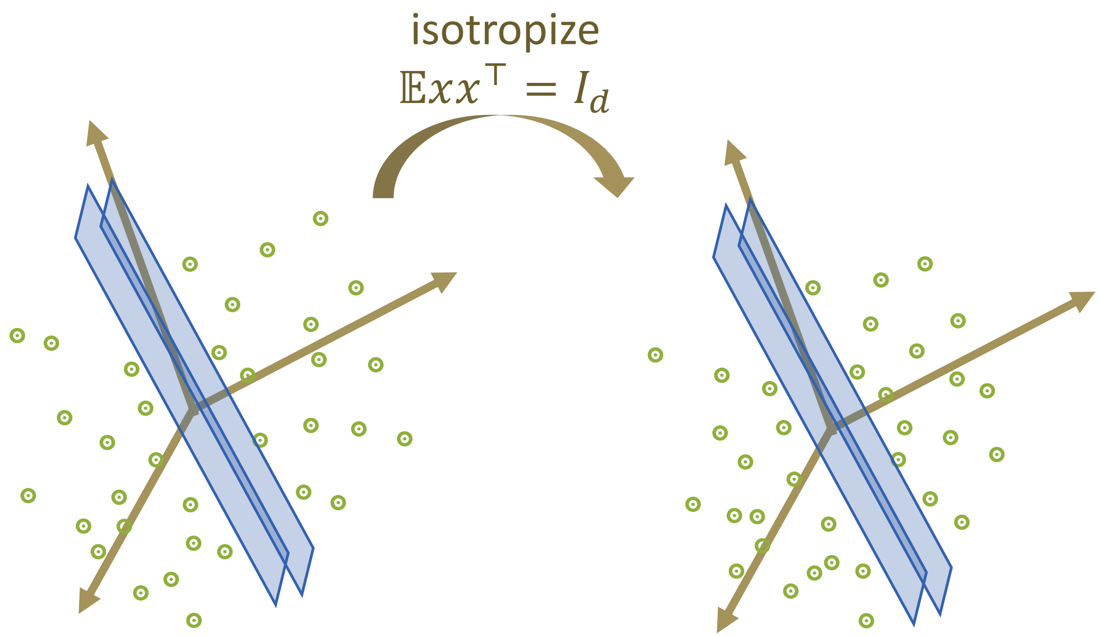
PCA!
(Principal Component Analysis)

Next: isotropize the data with margin



Next: isotropize the data with margin

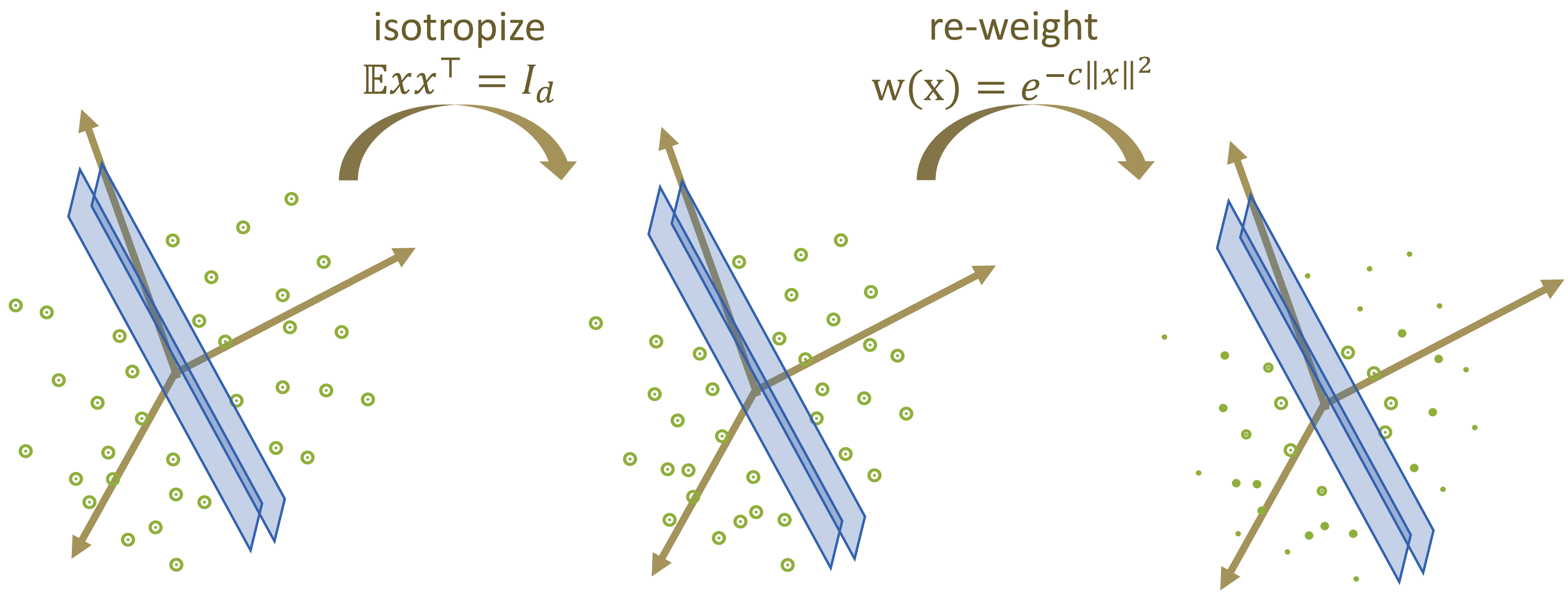
- Make the data isotropic (mean zero, covariance identity)
- PCA fails.



Input Data Distribution

Next: isotropize the data with margin

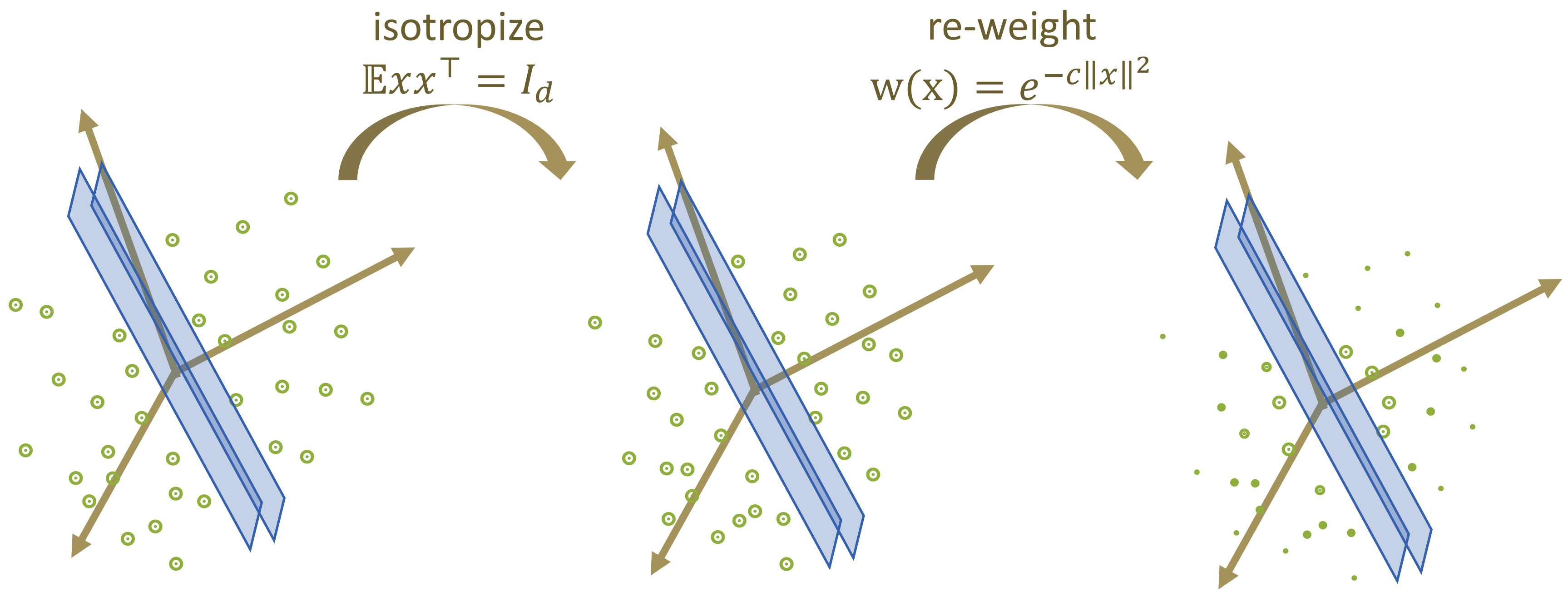
- Make the data isotropic (mean zero, covariance identity)
- PCA fails.
- Re-weight!



Input Data Distribution

Next: isotropize the data with margin

- Make the data isotropic (mean zero, covariance identity)
- PCA fails.
- Re-weight! ✓



Input Data Distribution

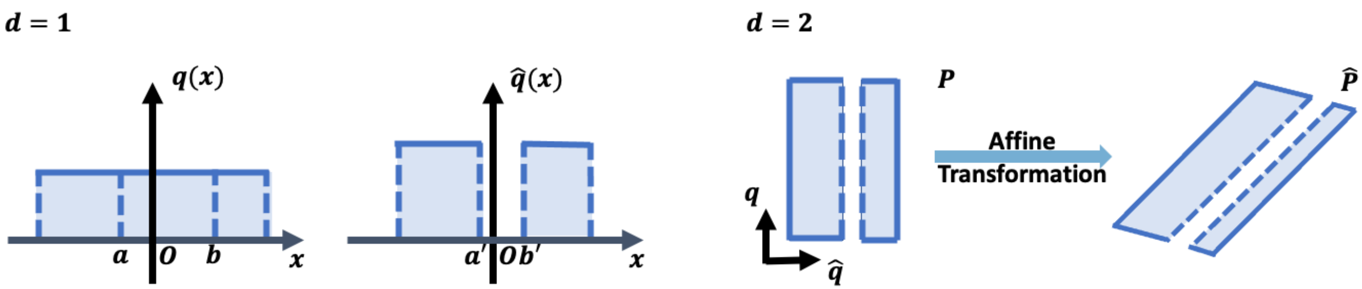
General data distribution

- No assumption of isotropy, or even mean zero!

General data distribution

- No assumption of isotropy, or even mean zero!

$\hat{P}$: **Affine transformation** of **product** of **symmetric logconcave** distributions with ϵ **margin** in an unknown direction u.



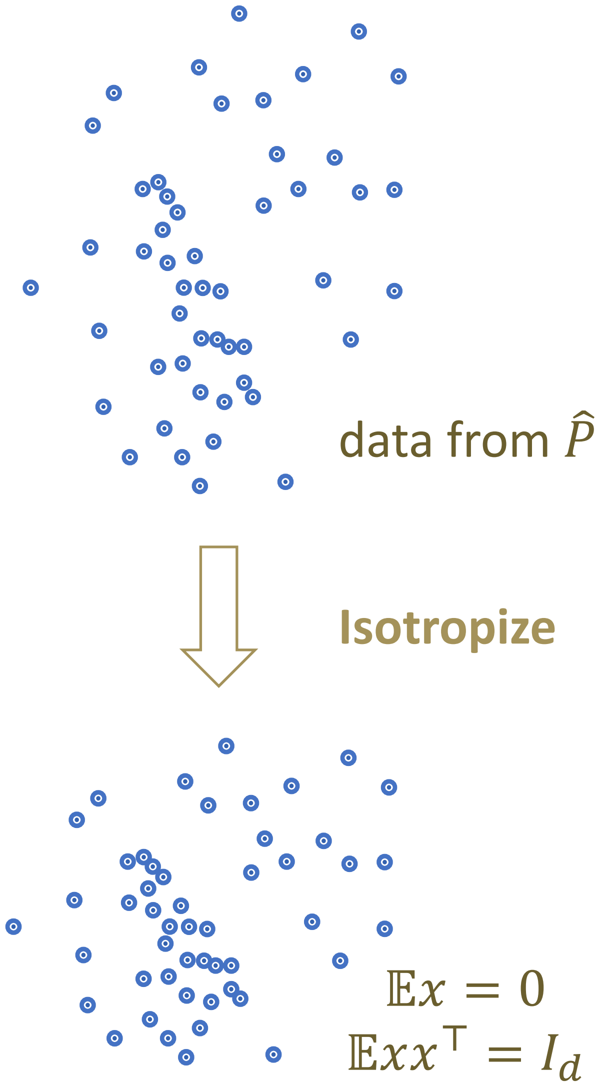
Known: data drawn from $\hat{P}$.
Unknown: logconcave distribution q, direction and location of the margin u, a, b, affine transformation A.

Algorithm: Contrastive Moments

Step 1: Make the data isotropic
(mean zero, covariance identity).

Step 2: Compute the **re-weighted sample mean** $\mu_i = \mathrm{Avg}\left(e^{\alpha_i \|x\|^2} x\right), i \in \{1,2\}$, and the **top eigenvector** v of **the re-weighted sample covariance** $\Sigma = \mathrm{Avg}(e^{\alpha_3 \|x\|^2} x x^\top)$.

Step 3: Project data along vectors μ_1, μ_2, v.
Output the one with the largest margin.

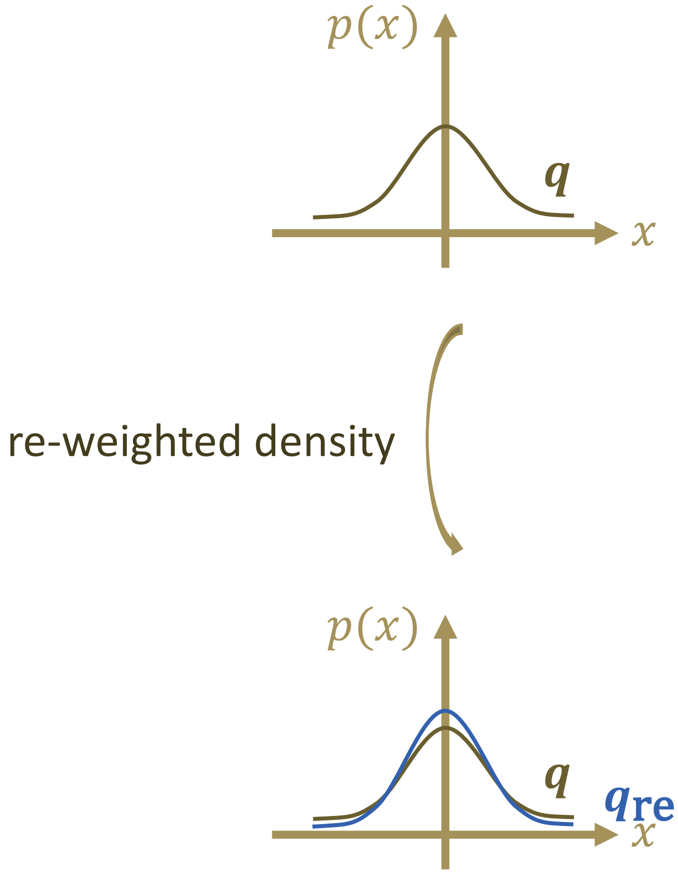


Algorithm: Contrastive Moments

Step 1: Make the data isotropic
(mean zero, covariance identity).

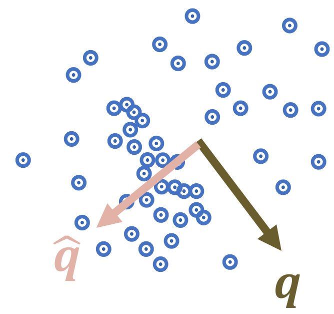
Step 2: Compute the **re-weighted sample mean** $\mu_i = \mathrm{Avg}\left(e^{\alpha_i \|x\|^2} x\right), i \in \{1,2\}$, and the **top eigenvector** v of **the re-weighted sample covariance** $\Sigma = \mathrm{Avg}(e^{\alpha_3 \|x\|^2} x x^\top)$.

Step 3: Project data along vectors μ_1, μ_2, v.
Output the one with the largest margin.

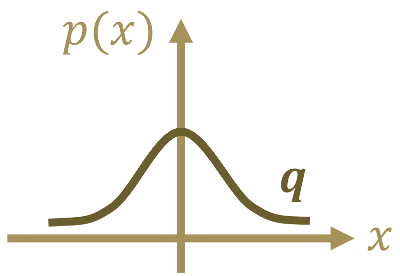


Algorithm: Contrastive Moments

Step 1: Make the data isotropic (mean zero, covariance identity).

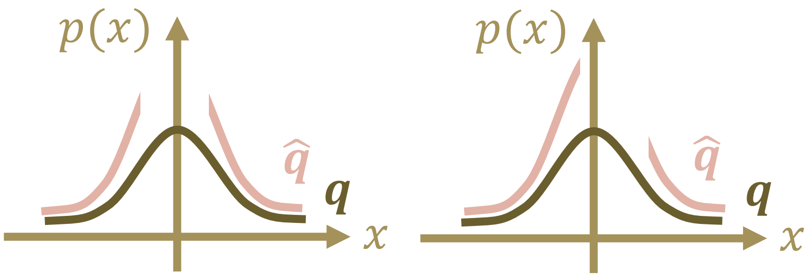


q: isotropic symmetric density
$\hat{q}$: isotropic density with margin



Step 2: Compute the **re-weighted sample mean** $\mu_i = \mathrm{Avg}\left(e^{\alpha_i \|x\|^2} x\right), i \in \{1,2\}$, and the **top eigenvector** v of **the re-weighted sample covariance** $\Sigma = \mathrm{Avg}(e^{\alpha_3 \|x\|^2} x x^\top)$.

Two cases of margin



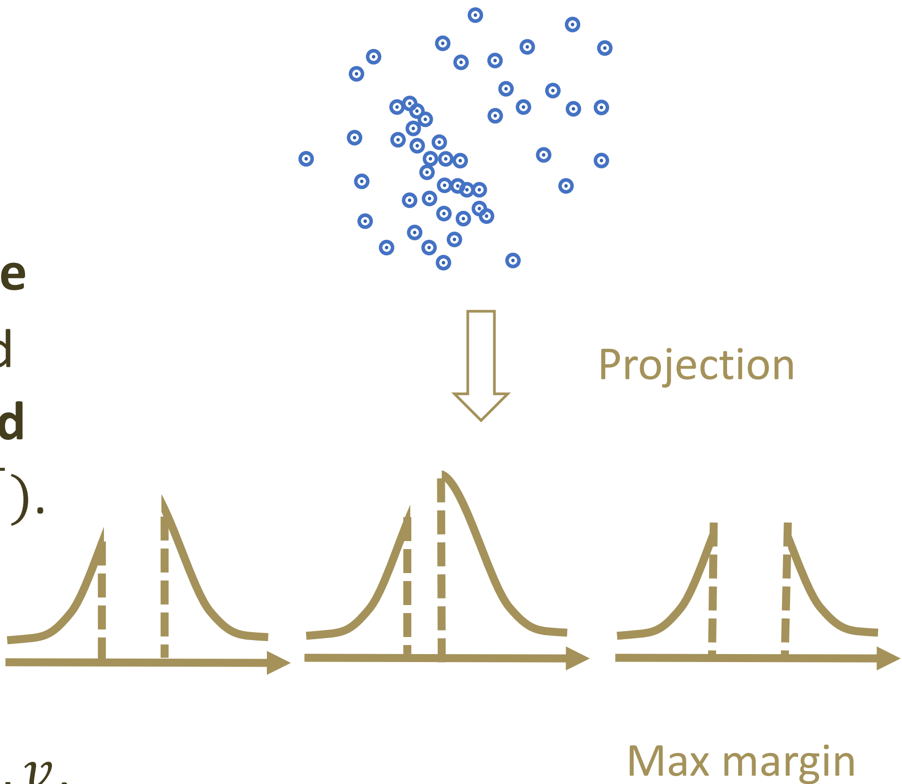
Step 3: Project data along vectors μ_1, μ_2, v. Output the one with the largest margin.

symmetric margin: re-weighted covariance
asymmetric margin: re-weighted mean

Algorithm: Contrastive Moments

Step 1: Make the data isotropic
(mean zero, covariance identity).

Step 2: Compute the **re-weighted sample mean** $\mu_i = \mathrm{Avg}\left(e^{\alpha_i \|x\|^2} x\right), i \in \{1,2\}$, and the **top eigenvector** v of **the re-weighted sample covariance** $\Sigma = \mathrm{Avg}(e^{\alpha_3 \|x\|^2} x x^\top)$.



Step 3: Project data along vectors μ_1, μ_2, v.
Output the one with the largest margin.

Main result

Theorem 1. There is an algorithm that can learn any **affine product logconcave distribution with ϵ-margin** to within TV distance δ with time and sample complexity that are $\mathbf{poly}(\boldsymbol{d}, \mathbf{1}/\boldsymbol{\varepsilon}, \mathbf{1}/\boldsymbol{\delta})$ with high probability.

Relevant Work

Non-gaussian Component Analysis (NGCA)

- Given distribution as a product of a d-1 dimensional Gaussian and a distinct distribution q in an unknown direction v.
- Goal: identify non-Gaussian direction v.
- Assumption: q and N(0,1) matches first k moments, and differs in k+1 moments.
- Order grows with k.
- To get ϵ TV distance, we need $k = \Omega\left(\log\left(\frac{1}{\epsilon}\right)\right)$.

Independent Component Analysis (ICA)

- Given samples from an unknown affine transformation of a product distribution.
- Goal: recover the affine transformation.
- Assumption: at most one component is Gaussian.

Future directions

- Analysis refinement.

  Linear in $d, 1/\epsilon$?
- Distribution generalization.
- Robust learning halfspaces.
- Intersection of halfspaces.
- Contrastive learning with data augmentation.

More in the paper :)