Sample Complexity of Learning Mahalanobis Distance Metrics

Nakul Verma
Janelia, HHMI

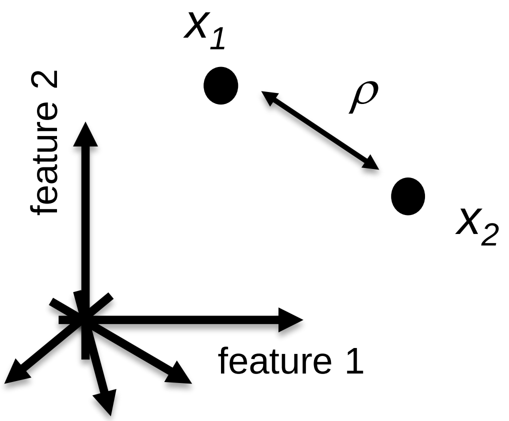
Mahalanobis Metric Learning

Comparing observations in feature space:

$$\rho(x_1, x_2) = \|x_1 - x_2\|^2$$
$$= (x_1 - x_2)^{\mathsf{T}}(x_1 - x_2)$$

[*sq.* ***Euclidean*** *dist*]

(all features are equally weighted)



$$\rho_M(x_1, x_2) = \|M(x_1 - x_2)\|^2$$

(using weighting mechanism M)

$$= (x_1 - x_2)^{\mathsf{T}}(M^{\mathsf{T}}M)(x_1 - x_2)$$

[*sq.* ***Mahalanobis*** *dist*]

Q: What should be the correct weighting *M*?

A: Data-driven.

Given data of interest, *learn* a *metric* (*M*), which helps in the prediction task.

Learning a Mahalanobis Metric

$$\rho_M(x_i, x_j) = (x_i - x_j)^\mathsf{T}(M^\mathsf{T}M)(x_i - x_j)$$

Suppose we want *M* s.t.:

- data from **same class** ≤ distance *U*
- data from **different classes** ≥ distance *L* [*U* < *L*]

Given two labelled samples $(x_i, y_i), (x_j, y_j)$ from a sample *S*. Then

- the distance between the pair $\rho_M^{ij} = \rho_M(x_i, x_j)$
- label agreement between the pair $Y_{ij} = \mathbf{1}[y_i = y_j]$

Define a **pairwise penalty** function
$$\phi(\rho_M^{ij}, Y_{ij}) = \begin{cases} (\rho_M^{ij} - U)_+ & \text{if } Y_{ij} = 1 \\ (L - \rho_M^{ij})_+ & \text{otherwise} \end{cases}$$

So **total** error:

$$\mathrm{err}_S(M) = \operatorname{avg}_{\substack{(x_i,y_i) \\ (x_j,y_j)}\in S} \left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

(empirical error over the sample S)

(error of M over the sample S)

$$\mathrm{err}(M) = \mathbb{E}\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

(generalization error)

(error of M over the (unseen) population)

Statistical consistency of Metric Learning

$$\mathrm{err}(M) = \mathbb{E}\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

Best possible metric on the population:

$$M^* = \mathrm{argmin}_M \ \ \mathrm{err}(M)$$

Best possible metric on the sample *S* (of size m) [drawn independently from the population]

$$M_m^* = \mathrm{argmin}_M \ \ \mathrm{err}_{S_m}(M)$$

Questions we want to answer:

(i) Does $\mathrm{err}(M_m^*) \to \mathrm{err}(M^*)$ as $m \to \infty$? (consistency)

(ii) At what rate does $\mathrm{err}(M_m^*) \to \mathrm{err}(M^*)$? (finite sample rates)

(iii) What factors affect the rate ? (data dim, feature info content)

What we show: Theorem 1

$$\mathrm{err}(M) = \mathbb{E}\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

Given a *D*-dimensional feature space.

For any λ-Lipschitz penalty function ϕ,

and any sample size *m*,

$$\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \le O\left(\lambda\sqrt{\frac{D\ln(1/\delta)}{m}}\right)$$

(with probability at least 1-δ over the draw of the sample)

If we want $\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \le \epsilon$,

then we require $m \ge \Omega\left(D\ln(1/\delta)\frac{\lambda^2}{\epsilon^2}\right)$

*This gives us **consistency** as well as a **rate**!*

Question: Is the convergence rate on the data dimension *D* tight?

What we show: Theorem 2

$$\mathrm{err}(M) = \mathbb{E}\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

Given a *D*-dimensional feature space.

For **any** metric learning algorithm *A* that (given a sample S_m) returns

$$A(S_m) = \mathrm{argmin}_M \ \mathrm{avg}_{S_m}\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

There exists a λ-Lipschitz penalty function ϕ, s.t. for all ε, δ,

if sample size $m \leq O(D/\epsilon^2)$

then

$$P_{S_m}\left[\mathrm{err}(A(S_m)) - \mathrm{err}(M^*) > \epsilon\right] > \delta$$

*Dependence on the representation dimension D is **tight**!*

***Remark**: this is the worst case analysis in the **absence** of any other information about the data distribution.*

*Can we **refine** our results if we know about the quality of our feature set?*

Quantifying feature-set quality

$$\mathrm{err}(M) = \mathbb{E}\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

Quantifying the quality of our feature set.

Observation: not all features are created equal

Each feature has a different information content for the prediction task.

Fix a particular prediction task *T*.

Let **M** be the optimal feature weighting for task *T*.

Define the *metric learning complexity* *d** for task *T* as: $d^* = \|\mathbf{M}^\mathsf{T}\mathbf{M}\|_F^2$

d is unknown a priori*

Question: Can we get a sample complexity rate that only depends on *d**?

What we show: Theorem 3

Given a *D*-dimensional feature space, and

a prediction task *T* with (unknown) metric learning complexity *d**

For any λ-Lipschitz penalty function ϕ ,

and any sample size *m,*

$$\mathrm{err}(M_m^{\mathrm{reg}}) - \mathrm{err}(M^*) \quad \leq O\left(\lambda\sqrt{\frac{d^* \ln(D)\ln(1/\delta)}{m}}\right)$$

(with probability at least 1-δ over the draw of the sample)

$$M_m^{\mathrm{reg}} = \mathrm{argmin}_M \quad \left[\mathrm{avg}_S\left[\phi(\rho_M^{ij}, Y_{ij})\right] + \Lambda\|M^\mathsf{T} M\|_{_F}\right] \qquad \Lambda \approx \lambda\sqrt{\ln(D/\delta)/m}$$

Take home message:

regularization can help adapt to the unknown metric learning complexity!

Empirical Evaluation

Want to study

Given a dataset with small metric learning complexity, but high representation dimension. How do regularized vs. unregularized Metric Learning algs. fare?

Approach

- pick benchmark datasets of low dimensionality (*d*)
- augment each dataset with large (*D* dim.) corr. noise

$$\Sigma_D \sim \text{Wishart}(\text{unit-scale})$$

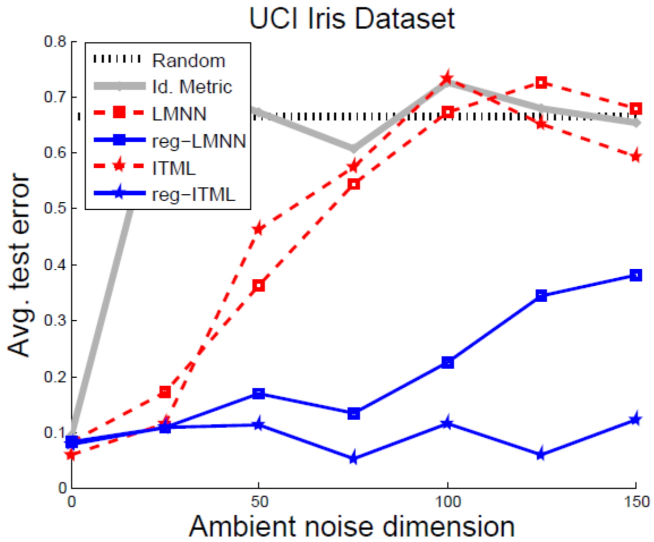
for each orig. sample x_i, augmented sample $x_i = [x_i \; x_\sigma]$ $\quad x_\sigma \sim N(0, \Sigma_D)$

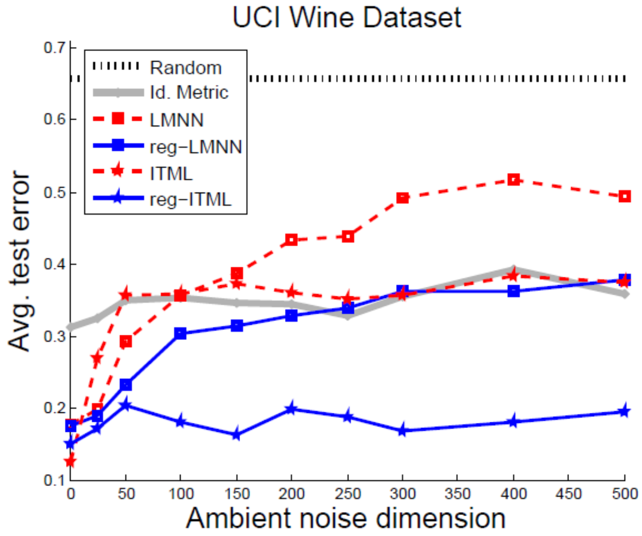
(we can now control signal-noise ratio)

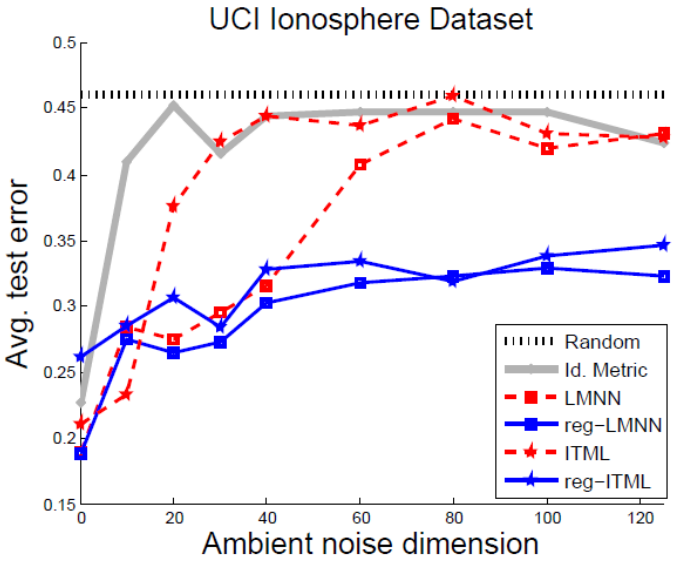
UCI dataset	dim (*d*)
Iris	4
Wine	13
Ionosphere	34

- study the prediction accuracy of regularized & unregularized Metric Learning algorithms as a function of noise dimension.

Empirical Evaluation







Theorem 1

$$\mathrm{err}(M) = \mathbb{E}\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S\left[\phi(\rho_M^{ij}, Y_{ij})\right]$$

Given a *D*-dimensional feature space.

For any λ-Lipschitz penalty function ϕ and any sample size *m*,

$$\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \leq O\left(\lambda\sqrt{\frac{D\ln(1/\delta)}{m}}\right)$$

(with probability at least 1-δ over the draw of the sample)

If we want $\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \leq \epsilon$,

then we require $m \geq \Omega\left(D\ln(1/\delta)\frac{\lambda^2}{\epsilon^2}\right)$

*This gives us **consistency** as well as a **rate**!*

How can we prove this?

Proof Idea (Theorem 1)

Want to find a sample size *m* such that for **any** weighting *M*

empirical performance of *M* ≈ generalization performance of *M*

Then, choosing the best M on samples, will have close to best generalization performance!

Try 1 (covering argument)

Fix a weighting metric *M*, define random variable

$$\mathrm{err}(M) = \mathbb{E}[\phi(\rho_M^{ij}, Y_{ij})]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S[\phi(\rho_M^{ij}, Y_{ij})]$$

$$\mathbf{Z}_{ij}^M = \phi(\rho_M^{ij}, Y_{ij}) \in [0,1]$$

if ϕ is bounded per example pair

By Hoeffding's bound (since ***Z*** *is bounded r.v.)*

$$\left|\mathrm{avg}_{S_m}[\mathbf{Z}_{ij}^M] - \mathbb{E}[\mathbf{Z}_{ij}^M]\right| \leq \sqrt{\frac{\ln(2/\delta)}{2m}}$$

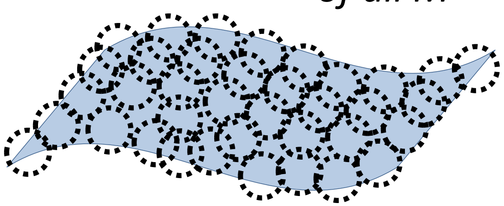
w.p. ≥ 1- δ over the draw of S_m

But, we want to have a similar result for all M!

For ***all*** *$M \in \mathcal{M}$*

$$\left|\mathrm{avg}_{S_m}[\mathbf{Z}_{ij}^M] - \mathbb{E}[\mathbf{Z}_{ij}^M]\right| \leq O\left(\sqrt{\frac{D^2 \ln(1/\delta)}{m}}\right)$$

$\mathcal{M}$ *Collection of all M*



Proof Idea (Theorem 1)

Try 2 (VC argument)

Recall: VC-theory is only for binary classification

If we view metric learning as classification, we can apply VC-style results!

Recall: given a (binary) classification class ***F***, for all $f \in \boldsymbol{F}$

$$\left|\text{avg}_{S_m}\left[\mathbf{1}[f(x_i) \neq y_i]\right] - \mathbb{E}\left[\mathbf{1}[f(x_i) \neq y_i]\right]\right| \leq O\left(\sqrt{\frac{\mathfrak{D}_F \ln(1/\delta)}{m}}\right)$$

w.p. $\geq 1-\delta$ *over the draw of* S_m

$\mathfrak{D}_F =$ *maximum sample size that can achieve all possible labels from using* $f \in \boldsymbol{F}$

For metric learning: say penalty function ϕ is binary threshold on distance.

$\boldsymbol{F} = \{ \phi_M : M \in \mathcal{M} \}$

+ labeling => ρ_M for a pair is small
– labeling => ρ_M for a pair is large

$$\text{err}(M) = \mathbb{E}[\phi(\rho_M^{ij}, Y_{ij})]$$
$$\text{err}_S(M) = \text{avg}_S[\phi(\rho_M^{ij}, Y_{ij})]$$

what is the maximum number of pairs which can attain all labeling from ***F*** ?

$$\mathfrak{D}_F \leq O(D^2)$$

(VC complexity of ellipsoids)

(a) only works for thresholds on ϕ
(b) cannot adapt to quality of the feature space!

Proof Idea (Theorem 1)

F = { ϕ_M : *M* $\in \mathcal{M}$}

Try 3 (Rademacher Complexity argument)

Rademacher Complexity: given a class ***F***, how well does some *f* ∈ ***F*** correlate to binary noise $\sigma \in \{-1, 1\}$.

$$\mathfrak{R}_F^m = \mathbb{E}_{x_i}\mathbb{E}_{\sigma_i}\sup_f\left|\frac{1}{m}\sum_i \sigma_i f(x_i)\right|$$

Then for all *f*

$$\mathbb{E}[f(x_i)] - \mathrm{avg}_{S_m}[f(x_i)] \quad \leq \quad 2\mathfrak{R}_F^m + O\left(\sqrt{\frac{\ln(1/\delta)}{m}}\right)$$

w.p. ≥ 1- δ over the draw of S_m

For metric learning

$$\mathfrak{R}_F^m \leq O\left(\sqrt{\frac{\sup_M \|M^\mathsf{T} M\|_F^2}{m}}\right)$$

(a) *works for any Lipschitz* ϕ

(b) ***can*** *adapt to quality of the feature space!*

for scale restricted metrics M, $\|M^T M\|^2 \leq D$

■

Theorem 2

$$\mathrm{err}(M) = \mathbb{E}\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

$$\mathrm{err}_S(M) = \mathrm{avg}_S\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

Given a *D*-dimensional feature space.

For **any** metric learning algorithm *A* that (given a sample S_m) returns

$$A(S_m) = \mathrm{argmin}_M \; \mathrm{avg}_{S_m}\big[\phi(\rho_M^{ij}, Y_{ij})\big]$$

There exists a λ-Lipschitz penalty function ϕ, s.t. for all ε, δ,

if sample size $m \leq O\big(D/\epsilon^2\big)$

then

$$P_{S_m}\big[\mathrm{err}(A(S_m)) - \mathrm{err}(M^*) > \epsilon\big] > \delta$$

Dependence on the representation dimension D is ***tight****!*

How can we prove this?

Proof Idea (Theorem 2)

Try 1: (VC argument, by treating Metric Learning as classification)

If we can lower bound $\mathfrak{D}_F \geq$ m ,

then a standard construction gives a specific distribution on which we must have $\Omega(m/\varepsilon^2)$ samples to get accuracy within ε.

Since, we work with pairs of points, the specific distribution for VC argument doesn't actually ever occur! (we need this distribution to be a product distribution)

Try 2: (Our approach -- deconstruct the VC argument)

We'll use the **probabilistic method**.

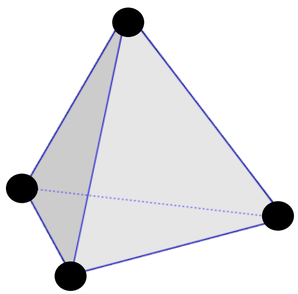
- Create a collection of distributions such that if one of them is chosen at random then the generalization error of *M* returned by *A* would be large.

So there is some distribution in the collection which has large error.

These distributions constructed so that Metric Learning acts as classification.

Proof Idea (Theorem 2)

Construction: (point masses on the vertices regular simplex)



- Collection of distributions:

  each vertex is labeled + or – (randomly) with bias ½ + ε
- Loss function:

$$\phi(\rho_M^{ij}, Y_{ij}) = \begin{cases} (\rho_M^{ij} - U)_+ & \text{if } Y_{ij} = 1 \\ (L - \rho_M^{ij})_+ & \text{otherwise} \end{cases}$$

[*U = 0, L = 1*]

Key insight: for this collection of distributions and this loss function the problem **reduces** to binary classification in the product space!

For *m* i.i.d. samples from a randomly selected dist. from the collection any empirical error minimizing algorithm would require $m \geq \Omega(D/\varepsilon^2)$

How? Calculate minimum number of samples required to distinguish the bias of two coins. Repeat it for ~D/2 pairs.

■

Other possible approaches:

Use information-theoretic arguments to establish minimum number of samples needed to distinguish good metric from bad ones. (e.g. use Fano's inequality)

Theorem 3

Given a *D*-dimensional feature space, and a prediction task *T* with (unknown) metric learning complexity *d**

For any λ-Lipschitz penalty function ϕ and any sample size *m,*

$$\mathrm{err}(M_m^{\mathrm{reg}}) - \mathrm{err}(M^*) \quad \leq O\left(\lambda\sqrt{\frac{d^* \ln(D)\ln(1/\delta)}{m}}\right)$$

(with probability at least 1-δ over the draw of the sample)

$$M_m^{\mathrm{reg}} = \mathrm{argmin}_M \quad \left[\mathrm{avg}_S\left[\phi(\rho_M^{ij}, Y_{ij})\right] + \Lambda\|M^\mathsf{T} M\|_{_F}\right] \qquad \Lambda \approx \lambda\sqrt{\ln(D/\delta)/m}$$

Take home message:

regularization can help adapt to the unknown metric learning complexity!

Proof Idea (Theorem 3)

Using Rademacher complexity argument, already shown:

$$\mathrm{err}(M_m^*) - \mathrm{err}(M^*) \;\leq O\left(\lambda \sqrt{\frac{\sup_M \|M^\mathsf{T} M\|_F^2 \cdot \ln(1/\delta)}{m}}\right)$$

$\leq D$

w.p. $\geq$ *1-* δ *over the draw of sample of size m*

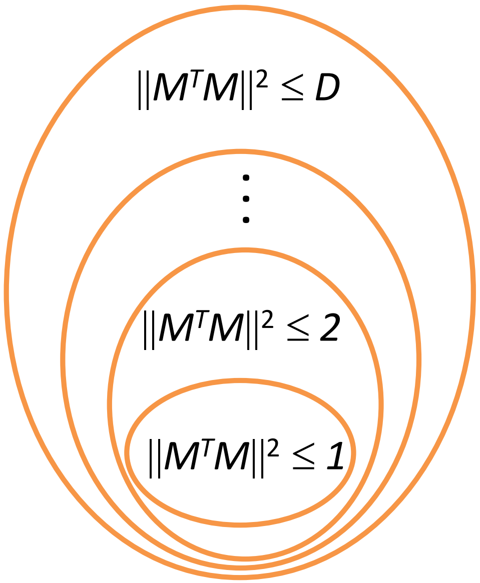
If we know *M** has small norm (say *d* << *D*), then we are done!

but don't know the norm of the best metric a priori…

Will use a refinement trick…

Observation: we are allowed to fail δ *fraction of time, we distribute this over each class* δ */ D*

A refinement of $\mathcal{M}$



For all $d \leq D$ and all M^d (s.t. $\|M^{d\,\mathsf{T}} M^d\|^2 \leq d$)

$$\mathrm{err}(M^d) - \mathrm{err}(M^d) \;\leq O\left(\lambda \sqrt{\frac{d \cdot \ln(D/\delta)}{m}}\right)$$

Proof Idea (Theorem 3)

$$\mathrm{err}(M^d) - \mathrm{err}(M^d) \ \leq O\left(\lambda\sqrt{\frac{d \cdot \ln(D/\delta)}{m}}\right)$$

So, if the algorithm picks:

$$M_m^{\mathrm{reg}} = \mathrm{argmin}_M \ \left[\mathrm{avg}_S\left[\phi(\rho_M^{ij}, Y_{ij})\right] + \Lambda \|M^{\mathsf{T}} M\|_F\right] \qquad \Lambda \approx \lambda\sqrt{\ln(D/\delta)/m}$$

Then (w.p. ≥ 1- δ):

$$\begin{aligned}
\mathrm{err}(M_m^{\mathrm{reg}}) - \mathrm{err}(M^*) \quad & \leq \mathrm{err}_{S_m}(M_m^{\mathrm{reg}}) + \Lambda \|(M_m^{\mathrm{reg}})^{\mathsf{T}}(M_m^{\mathrm{reg}})\|_F - \mathrm{err}(M^*) \\
& \leq \mathrm{err}_{S_m}(M^*) + \Lambda \|(M^*)^{\mathsf{T}}(M^*)\|_F - \mathrm{err}(M^*) \\
& = O\left(\lambda\sqrt{\frac{d^* \ln(D/\delta)}{m}}\right)
\end{aligned}$$

■

Comparison with previous results

	Previous results	Our results
Convergence rate (upper bound)	For thresholds on convex ϕ $\leq O(\sqrt{?/m})$ Stable and regularized algs. $\leq O(\sqrt{1/m})$	For general Lipschitz ϕ with ERM $\leq O(\sqrt{D/m})$ Theorem 1
Convergence rate (lower bound)	No known results	In absence of any other information, exists Lipschitz ϕ, with ERM $\geq \Omega(\sqrt{D/m})$ Theorem 2
Data complexity *d**	No known results	For gen. Lipschitz ϕ with regularized ERM $\leq O(\sqrt{d^* \ln(D)/m})$ Theorem 3

Open problems

- Analysis of Metric Learning in Online and Active Learning framework?
- Non-linear metric learning?
- ‘Structured’ metric learning? (ranking problems, clustering problems, etc)

Questions / Discussion

Thank You!