Mirror Descent for Metric Learning

Gautam Kunapuli

Jude W. Shavlik

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

And what do we have here?

We have a ***metric learning algorithm*** *that uses* ***composite mirror descent*** *(COMID)*:

- **Unifying framework for metric learning.**
  - *Different algorithms from various Bregman and loss functions.*
- **Sparse metric.**
  - *Uses* ***trace-norm*** *regularization. This ensures that learned metric* ***is sparse*** *in its eigen-spectrum; only $r < n$ EVs used*
- **Scalability.**
  - *Updates require rank-1 modification of the EVD at each iteration;* ***implemented efficiently*** *and* ***embarrassingly parallel****.*
- **Kernelizable.**

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Problem Formulation

Learn a **pseudo-metric**

$$d_M(\mathbf{x}, \mathbf{z})^2 = (\mathbf{x}-\mathbf{z})' L' L (\mathbf{x}-\mathbf{z}) = (\mathbf{x}-\mathbf{z})' M (\mathbf{x}-\mathbf{z})$$

from **pairs of labeled data** points, $(\mathbf{x}_t, \mathbf{z}_t, y_t)_{t=1}^{T}$, where label y_t denotes **similarity/dissimilarity**

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Problem Formulation

- The following constraints should hold

$$\forall(\mathbf{x}, \mathbf{z}, y = +1) \;\Rightarrow\; d_M(\mathbf{x}, \mathbf{z})^2 \leq \mu - 1,$$
$$\forall(\mathbf{x}, \mathbf{z}, y = -1) \;\Rightarrow\; d_M(\mathbf{x}, \mathbf{z})^2 \geq \mu + 1,$$

such that **similar points** are **transformed closer** together, while **dissimilar points** are **transformed farther** apart under *L*:

$$d(\mathbf{x}, \mathbf{z}) = \|L(\mathbf{x} - \mathbf{z})\|_2$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Problem Formulation

- The following constraints

$$\forall(\mathbf{x}, \mathbf{z}, y = +1) \;\Rightarrow\; d_M(\mathbf{x}, \mathbf{z})^2 \;\leq\; \mu - 1,$$
$$\forall(\mathbf{x}, \mathbf{z}, y = -1) \;\Rightarrow\; d_M(\mathbf{x}, \mathbf{z})^2 \;\geq\; \mu + 1,$$

can be **rewritten** compactly as

$$y(\mu - d_M(\mathbf{x}, \mathbf{z})^2) \;\geq\; 1$$

$$d_M(\mathbf{x}, \mathbf{z})^2 \;=\; (\mathbf{x} - \mathbf{z})' M (\mathbf{x} - \mathbf{z})$$

Problem Formulation

- The following constraints

$$\forall(\mathbf{x}, \mathbf{z}, y = +1) \ \Rightarrow \ d_M(\mathbf{x},\, \mathbf{z})^2 \ \le \ \mu - 1,$$
$$\forall(\mathbf{x}, \mathbf{z}, y = -1) \ \Rightarrow \ d_M(\mathbf{x},\, \mathbf{z})^2 \ \ge \ \mu + 1,$$

can be **rewritten** compactly as

$$y(\mu - d_M(\mathbf{x},\, \mathbf{z})^2) \ge 1$$

$$d_M(\mathbf{x},\, \mathbf{z})^2 \ = \ (\mathbf{x} - \mathbf{z})' M (\mathbf{x} - \mathbf{z})$$

this is the **margin function**, which can be used to define several different loss functions

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Problem Formulation

- The following constraints

$$\forall(\mathbf{x}, \mathbf{z}, y = +1) \;\Rightarrow\; d_M(\mathbf{x},\, \mathbf{z})^2 \;\leq\; \mu - 1,$$
$$\forall(\mathbf{x}, \mathbf{z}, y = -1) \;\Rightarrow\; d_M(\mathbf{x},\, \mathbf{z})^2 \;\geq\; \mu + 1,$$

can be **rewritten** compactly as

$$y(\mu - d_M(\mathbf{x},\, \mathbf{z})^2) \geq 1$$

$$d_M(\mathbf{x},\, \mathbf{z})^2 \;=\; (\mathbf{x} - \mathbf{z})' M (\mathbf{x} - \mathbf{z})$$

this is the **margin function**,
$m(M, \mu; \mathbf{x}, \mathbf{z}, y)$

For instance: the **hinge loss**

$$\ell(M, \mu) = \max\{\, 0,\, 1 - m(M, \mu; \mathbf{x}, \mathbf{z}, y)\,\}$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Outline

- Introduction
- **Mirror Descent for Metric Learning**
  - Formulation
  - Loss Functions and Bregman Functions
  - Closed-form Updates
  - Efficient Implementation
- Experiments
  - Results: Benchmark Data Sets
  - Results: OptDigits Data Set
- Conclusions

Mirror Descent

- Mirror descent (MD; Beck & Teboulle, 2003) is a **proximal-gradient method** for minimizing a convex function,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} B_\psi(\mathbf{w}, \mathbf{w}_t) + \eta\nabla'\phi_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Mirror Descent

- Mirror descent (MD; Beck & Teboulle, 2003) is a **proximal-gradient method** for minimizing a convex function,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} B_{\psi}(\mathbf{w},\, \mathbf{w}_t) + \eta\, \nabla'\phi_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)$$

Bregman function, to measure proximity between iterates

Gradient of the convex function

Mirror Descent

- Mirror descent (MD; Beck & Teboulle, 2003) is a **proximal-gradient method** for minimizing a convex function,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} B_\psi(\mathbf{w}, \mathbf{w}_t) + \eta \nabla' \phi_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)$$

- Composite mirror descent (COMID; Duchi et al, 2010) **generalizes MD** to loss-and-regularization composite functions $\phi_t = \ell_t + r$

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} B_\psi(\mathbf{w}, \mathbf{w}_t) + \eta \nabla' \ell_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) + \eta\, r(\mathbf{w})$$

Mirror Descent

- Mirror descent (MD; Beck & Teboulle, 2003) is a proximal-gradient method for minimizing a convex function,

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} B_\psi(\mathbf{w}, \mathbf{w}_t) + \eta \nabla' \phi_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t)$$

- Composite mirror descent (COMID; Duchi et al, 2010) generalizes MD to loss-and-regularization composite functions $\phi_t = \ell_t + r$

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\Omega} B_\psi(\mathbf{w}, \mathbf{w}_t) + \eta \nabla' \ell_t(\mathbf{w}_t)(\mathbf{w} - \mathbf{w}_t) + \eta\, r(\mathbf{w})$$

only **loss is linearized**; **regularization** is **not linearized**

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Mirror Descent For Metric Learning

- Learn pseudo-metric incrementally from triplets, and at each iteration, compute updates:

$$M_{t+1} = \underset{M \succeq 0}{\arg\min}\; B_\psi(M, M_t) + \eta \, \langle \nabla_M \ell_t(M_t, \mu_t), \, M - M_t \rangle + \eta \, \rho \; |||\, M \,|||$$

$$\mu_{t+1} = \underset{\mu \geq 1}{\arg\min} \quad B_\psi(\mu, \mu_t) + \eta \, \nabla_\mu \ell_t(M_t, \mu_t)' \; (\mu - \mu_t).$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Mirror Descent For Metric Learning

- Learn pseudo-metric incrementally from triplets, and at each iteration, compute updates:

$$M_{t+1} = \arg\min_{M \succeq 0} B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \rho \, ||| M |||$$

$$\mu_{t+1} = \arg\min_{\mu \geq 1} \quad B_\psi(\mu, \mu_t) + \eta \nabla_\mu \ell_t(M_t, \mu_t)' \, (\mu - \mu_t).$$

metric matrix should be symmetric, **positive semidefinite**

margin should be at least 1 to ensure that **learned distance is positive**

Mirror Descent For Metric Learning

- Learn pseudo-metric incrementally from triplets, and at each iteration, compute updates:

$$M_{t+1} = \arg\min_{M \succeq 0} B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \rho \, |||M|||$$

$$\mu_{t+1} = \arg\min_{\mu \geq 1} \quad B_\psi(\mu, \mu_t) + \eta \nabla_\mu \ell_t(M_t, \mu_t)' \, (\mu - \mu_t).$$

various loss and Bregman functions can be used to derive **different classes of algorithms**

Mirror Descent For Metric Learning

- Learn pseudo-metric incrementally from triplets, and at each iteration, compute updates:

$$M_{t+1} = \arg\min_{M \succeq 0} B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \rho \, ||| M |||$$

$$\mu_{t+1} = \arg\min_{\mu \geq 1} \quad B_\psi(\mu, \mu_t) + \eta \nabla_\mu \ell_t(M_t, \mu_t)' \, (\mu - \mu_t).$$

The trace norm is the **sum of the singular values of a matrix**,

$$|||X||| = \mathbf{e}' |\boldsymbol{\lambda}|$$

trace-norm regularization is used to produce a metric that is **sparse in its eigenspectrum**

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Loss Functions

- Some (Lipschitz) loss functions for metric learning, where the margin function is $m_t(\mathbf{u}_t, y_t) = y_t(\mu - \mathrm{tr}\, M \mathbf{u}_t \mathbf{u}_t')$ and $\mathbf{u}_t = \mathbf{x}_t - \mathbf{z}_t$

Loss	$\ell_t(M_t, \mu_t)$	$\nabla_M \ell_t(M_t, \mu_t)$
Hinge	$(1 - m_t)_+$	$(1 - m_t)_\star \; (y_t \mathbf{u}_t \mathbf{u}_t')$
Modified Least Sq.	$\frac{1}{2}(1 - m_t)_+^2$	$(1 - m_t)_+ \, (y_t \mathbf{u}_t \mathbf{u}_t')$
Logistic	$\log\left(1 + \exp(-m_t)\right)$	$\frac{\exp(-m_t)}{1+\exp(-m_t)} (y_t \mathbf{u}_t \mathbf{u}_t')$

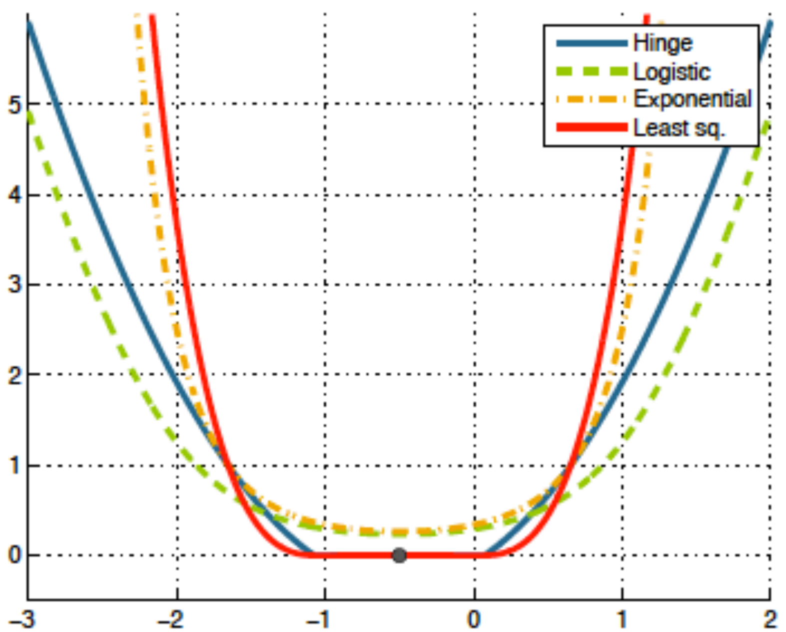
$()_+$ is the max function

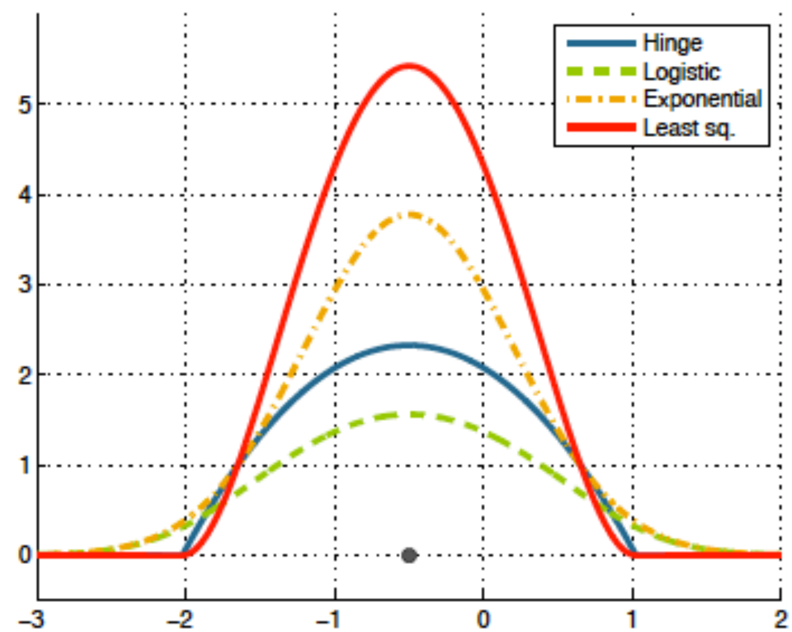
$()_*$ is the step function

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Loss Functions

- Behavior of various loss functions around x = *-0.5*, when **(left)** with **similar points** and pair labels: y = 1, and **(right)** with **dissimilar points** and pair labels, *y = -1*



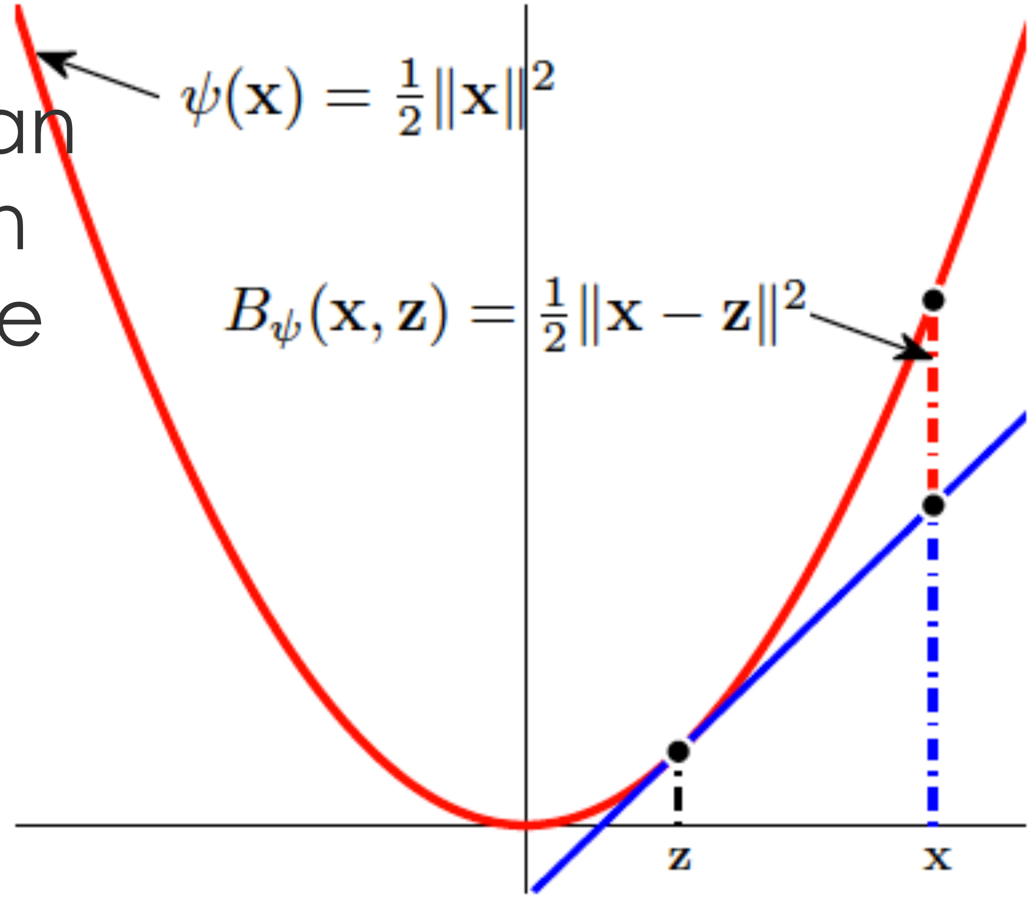


WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Bregman Functions

Squared Euclidean distance is a Bregman divergence and can be generalized in the **matrix case** to the **squared Frobenius distance:**

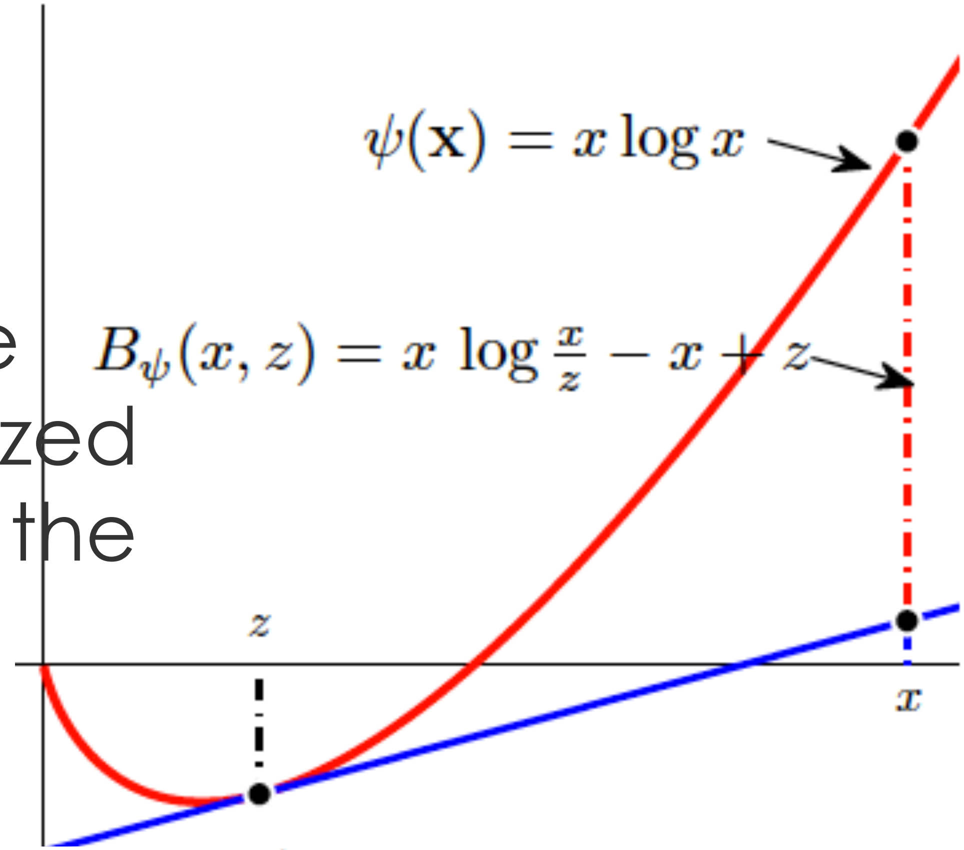
$$B_\psi(X,Z) = \frac{1}{2}\|X - Z\|_F^2$$



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Bregman Functions

Kullback-Liebler (KL) divergence is a Bregman divergence and can be generalized in the **matrix case** to the **von Neumann divergence:**



$$B_\psi(X, Y) = \mathrm{tr}\left(X \log X - X \log Y - X + Y\right)$$

Generalized Update Rules

$$M_{t+1} = \underset{M \succeq 0}{\arg\min}\ B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \rho \, |||M|||$$

For general choice of Bregman function and loss, update rules can be derived in closed-form using the **eigenvalue thresholding (shrinkage) operator**

$$S_\tau(X) = V \operatorname{diag}(\boldsymbol{\lambda}_\tau) V'$$

$$(\lambda_\tau)_i = (\lambda_i - \tau)_+$$

which cuts off all eigenvalues below the specified threshold, τ

Generalized Update Rules

$$M_{t+1} = \arg\min_{M \succeq 0} \; B_\psi(M, M_t) + \eta \, \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \, \rho \; |||M|||$$

Update rules can be derived in closed-form using the **eigenvalue thresholding/shrinkage operator**: $S_\tau(X) = V \operatorname{diag}(\boldsymbol{\lambda}_\tau) V'$, where $(\lambda_\tau)_i = \operatorname{sign}(\lambda_i) \max\{|\lambda_i| - \tau, \; \}$. The closed-form solutions are:

$$\textbf{vonNeumann} \quad M_{t+1} = \exp\left(S_{\eta\rho}(\log M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right),$$

$$\textbf{Frobenius} \quad M_{t+1} = S_{\eta\rho}\left(M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right).$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Generalized Update Rules

$$M_{t+1} = \arg\min_{M \succeq 0} \; B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta\, \rho \; |||\, M \,|||$$

Update rules can be derived in closed-form using the **eigenvalue thresholding/shrinkage operator**: $S_\tau(X) = V \operatorname{diag}(\boldsymbol{\lambda}_\tau) V'$, where $(\lambda_\tau)_i = \operatorname{sign}(\lambda_i) \max\{|\lambda_i| - \tau, \ \}$. The closed-form solutions are:

$$\begin{aligned} \textbf{vonNeumann} \quad M_{t+1} &= \exp\left(S_{\eta\rho}(\log M_t - \eta \nabla_M \ell_t(M_t, \mu_t)) \right), \\ \textbf{Frobenius} \quad M_{t+1} &= S_{\eta\rho}(M_t - \eta \nabla_M \ell_t(M_t, \mu_t)) \,. \end{aligned}$$

eigenvalues are thresholded by **learning rate** (η) and the **regularization parameter** (ρ)

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Generalized Update Rules

$$M_{t+1} = \arg\min_{M \succeq 0} \; B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \rho \; ||| M |||$$

Update rules can be derived in closed-form using the **eigenvalue thresholding/shrinkage operator**: $S_\tau(X) = V \operatorname{diag}(\boldsymbol{\lambda}_\tau) V'$, where $(\lambda_\tau)_i = \operatorname{sign}(\lambda_i) \max\{|\lambda_i| - \tau, \}$. The closed-form solutions are:

$$\begin{aligned} \textbf{vonNeumann} \quad & M_{t+1} = \exp\left(S_{\eta\rho}(\log M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right), \\ \textbf{Frobenius} \quad & M_{t+1} = S_{\eta\rho}\left(M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right). \end{aligned}$$

For von Neumann divergence, note that exp is applied after thresholding: **smallest eigen-value is 1, not zero**.

Final **learned metric matrix is of full-rank**. However, can still **perform feature selection by dropping k smallest** eigen-values similar to PCA.

Generalized Update Rules

$$M_{t+1} = \arg\min_{M \succeq 0} \; B_\psi(M,\, M_t) + \eta \,\langle \nabla_M \ell_t(M_t, \mu_t),\, M - M_t \rangle + \eta\, \rho \; ||| \, M ||| $$

Update rules can be derived in closed-form using the **eigenvalue thresholding/shrinkage operator**: $S_\tau(X) = V \,\mathrm{diag}(\boldsymbol{\lambda}_\tau)\, V'$, where $(\lambda_\tau)_i = \mathrm{sign}(\lambda_i)\, \max\{|\lambda_i| - \tau, \;\}$. The closed-form solutions are:

$$\textbf{vonNeumann} \quad M_{t+1} = \exp\left(S_{\eta\rho}(\log M_t - \eta \nabla_M \ell_t(M_t, \mu_t)) \right),$$

$$\textbf{Frobenius} \quad M_{t+1} = S_{\eta\rho}\left(M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right).$$

Loss	$\ell_t(M_t,\, \mu_t)$	$\nabla_M\, \ell_t(M_t,\, \mu_t)$
Hinge	$(1 - m_t)_+$	$(1 - m_t)_\star \; (y_t \mathbf{u}_t \mathbf{u}_t')$
Modified Least Sq.	$\frac{1}{2}(1 - m_t)_+^2$	$(1 - m_t)_+ \, (y_t \mathbf{u}_t \mathbf{u}_t')$
Logistic	$\log\left(1 + \exp(-m_t)\right)$	$\frac{\exp(-m_t)}{1 + \exp(-m_t)} \, (y_t \mathbf{u}_t \mathbf{u}_t')$

gradients of the loss function are generally of the form

$$\nabla_M \ell_t = \alpha_t \mathbf{u}_t \mathbf{u}_t'$$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Generalized Update Rules

$$M_{t+1} = \arg\min_{M \succeq 0} \; B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta \rho \; \| M \|$$

Update rules can be derived in closed-form using the **eigenvalue thresholding/shrinkage operator**: $S_\tau(X) = V \operatorname{diag}(\boldsymbol{\lambda}_\tau) V'$, where $(\lambda_\tau)_i = \operatorname{sign}(\lambda_i) \max\{|\lambda_i| - \tau, \}$. The closed-form solutions are:

$$\textbf{vonNeumann} \quad M_{t+1} = \exp\left(S_{\eta\rho}(\log M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right),$$

$$\textbf{Frobenius} \quad M_{t+1} = S_{\eta\rho}\left(M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right).$$

At the t-th iteration, with $M_t = V_t \nabla \psi(\Lambda_t) V_t'$, we have:

(Intermediate gradient) $\quad \nabla \psi(M_{t+\frac{1}{2}}) = V_t \nabla \psi(\Lambda_t) V_t' - \alpha \mathbf{u}_t \mathbf{u}_t'$

(EVD of intermediate gradient) $\quad \nabla \psi(M_{t+\frac{1}{2}}) = V_{t+1} \Lambda_{t+1} V_{t+1}'$

(Matrix update/thresholding) $\quad M_{t+1} = V_{t+1} \nabla \psi^{-1}\left(S_{\eta\rho}(\Lambda_{t+1}) \right) V_{t+1}'$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Generalized Update Rules

$$M_{t+1} = \underset{M \succeq 0}{\arg\min}\; B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta\,\rho\, |||\, M |||$$

*update simply requires **rank-one modification** of current eigen-decomposition, followed by **thresholding of eigen-values!***

Update rules can be derived in closed-form using the **eigenvalue thresholding/shrinkage operator**: $S_\tau(X) = V\,\mathrm{diag}(\boldsymbol{\lambda}_\tau)\,V'$, where $(\lambda_\tau)_i = \mathrm{sign}(\lambda_i)\,\max\{|\lambda_i| - \tau,\ \}$. The closed-form solutions are:

$$\textbf{vonNeumann}\quad M_{t+1} = \exp\left(S_{\eta\rho}(\log M_t - \eta \nabla_M \ell_t(M_t, \mu_t)) \right),$$

$$\textbf{Frobenius}\quad M_{t+1} = S_{\eta\rho}\left(M_t - \eta \nabla_M \ell_t(M_t, \mu_t) \right).$$

At the t-th iteration, with $M_t = V_t \nabla\psi(\Lambda_t) V_t'$, we have:

(Intermediate gradient) $\quad \nabla\psi(M_{t+\frac{1}{2}}) = V_t \nabla\psi(\Lambda_t) V_t' - \alpha \mathbf{u}_t \mathbf{u}_t'$

(EVD of intermediate gradient) $\quad \nabla\psi(M_{t+\frac{1}{2}}) = V_{t+1}\, \Lambda_{t+1}\, V_{t+1}'$

(Matrix update/thresholding) $\quad M_{t+1} = V_{t+1}\, \nabla\psi^{-1}\left(S_{\eta\rho}(\Lambda_{t+1}) \right) V_{t+1}'$

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Efficient Implementation of Rank-One EVD Updates

A general update

$$M_{t+1} = V_t \nabla \psi(\Lambda_t) V_t' - \alpha \mathbf{u}_t \mathbf{u}_t'$$

involves a **rank-one modification of the EVD** at the current iteration

It is known that the eigen-values of the two matrices **interlace**

Each **new eigen-value can be computed independently**, as it is bounded between two old eigen-values

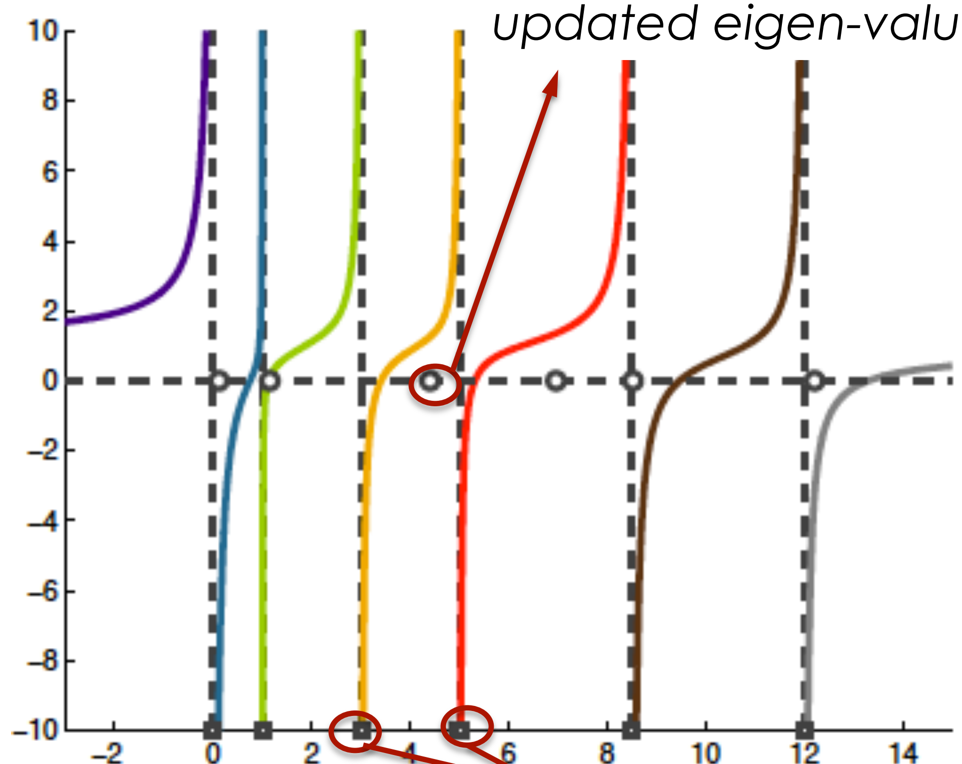


Figure: *Plot of the* ***secular equation*** *of the rank-one perturbation*

Efficient Implementation of Rank-One EVD Updates

- In general, **any root-finding technique** (eg., Newton-Raphson) can be used to compute eigen-values independently from the secular equation

- May result in **non-orthogonal eigen-vectors**. Instead, we implement **rational interpolation approach of Gu and Eisenstat** (1994)

- **Efficiency** of approach **increases as multiplicity of** repeated EVs **increases**

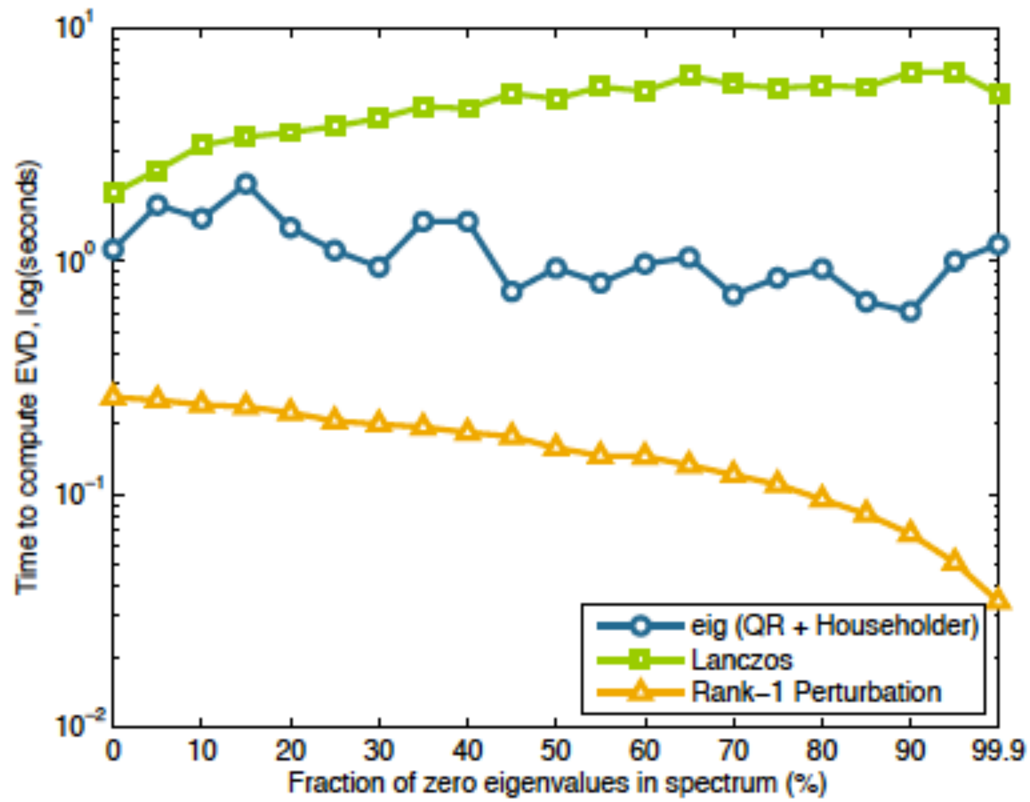


Figure: *Comparing various eigen-value decomposition algorithms with the rank-one perturbation approach*

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Mirror Descent for Metric Learning

input: data $(\mathbf{x}_t, \mathbf{z}_t, y_t)_{t=1}^{T}$, parameters $\rho,\ \eta > 0$
choose: Bregman functions $\psi(M)$; $\psi(\mu)$, loss $\ell(M, \mu)$
initialize: $M_0 = I_n$, $\mu_0 = 1$
for $(\mathbf{x}^t,\ \mathbf{z}_t,\ y_t)$ **do**
    let $\mathbf{u}_t = \mathbf{x}_t - \mathbf{z}_t$, $\eta_t = \eta/\sqrt{t}$
    compute gradients of loss $\nabla_M \ell_t = \alpha_t \mathbf{u}_t \mathbf{u}_t'$ and $\nabla_\mu \ell_t = -\alpha_t$
    write $\nabla\psi(M_t) = V_t \nabla\psi(\Lambda_t) V_t'$
    rank-one update $V_{t+1}\,\Lambda_{t+1}\,V_{t+1}' = V_t \nabla\psi(\Lambda_t) V_t' - \alpha \mathbf{u}_t \mathbf{u}_t'$
    shrink the eigenvalues $M_{t+1} = V_{t+1} \nabla\psi^{-1}\left(S_{\eta\rho}(\Lambda_{t+1}) \right) V_{t+1}'$
    margin update $\mu_{t+1} = \max\left(\nabla\psi^{-1}\left(\nabla\psi(\mu_t) - \eta\,\nabla\ell_t(M_t, \mu_t)\right),\ 1\right)$
end for

Outline

- Introduction
- Mirror Descent for Metric Learning
  - Formulation
  - Loss Functions and Bregman Functions
  - Closed-form Updates
  - Efficient Implementation
- **Experiments**
  - Results: Benchmark Data Sets
  - Results: OptDigits Data Set
- Conclusions

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Benchmark Data Sets

- We consider two algorithms
  - MDML: **Frobenius div. and hinge loss** *(MDML H+F)*
  - MDML: **von Neumann div. and log. loss** *(MDML L+V)*
- We compare these approaches to four well-known batch and online metric learning approaches
  - **large-margin nearest neighbor** *(Weinberger et al, 2006)*
  - **information-theoretic metric learning** ***(Davis et al, 2007)***
  - **BoostMetric** *(Shen et al, 2009)*
  - **pseudo-metric online learning** *(Shalev-Shwartz et al, 2004)*

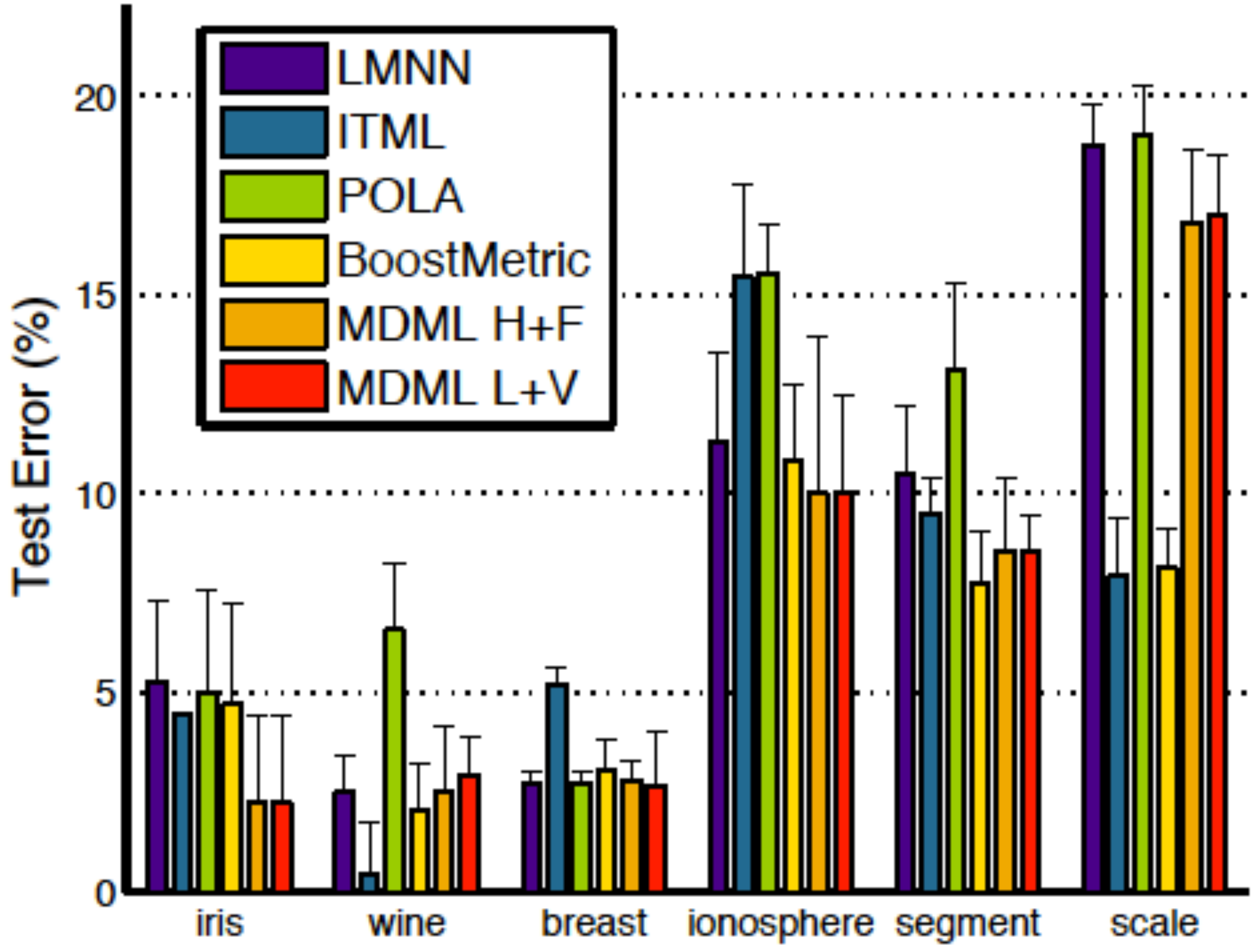
WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Benchmark Data Sets

- Triplets for learning **generated** using the same strategy as **Weinberger et al** (2006)
  - For each training point *k*=3 similarly labeled (targets) and *k*=3 differently labeled (impostors) are selected
  - Test data classified using 3-NN classification

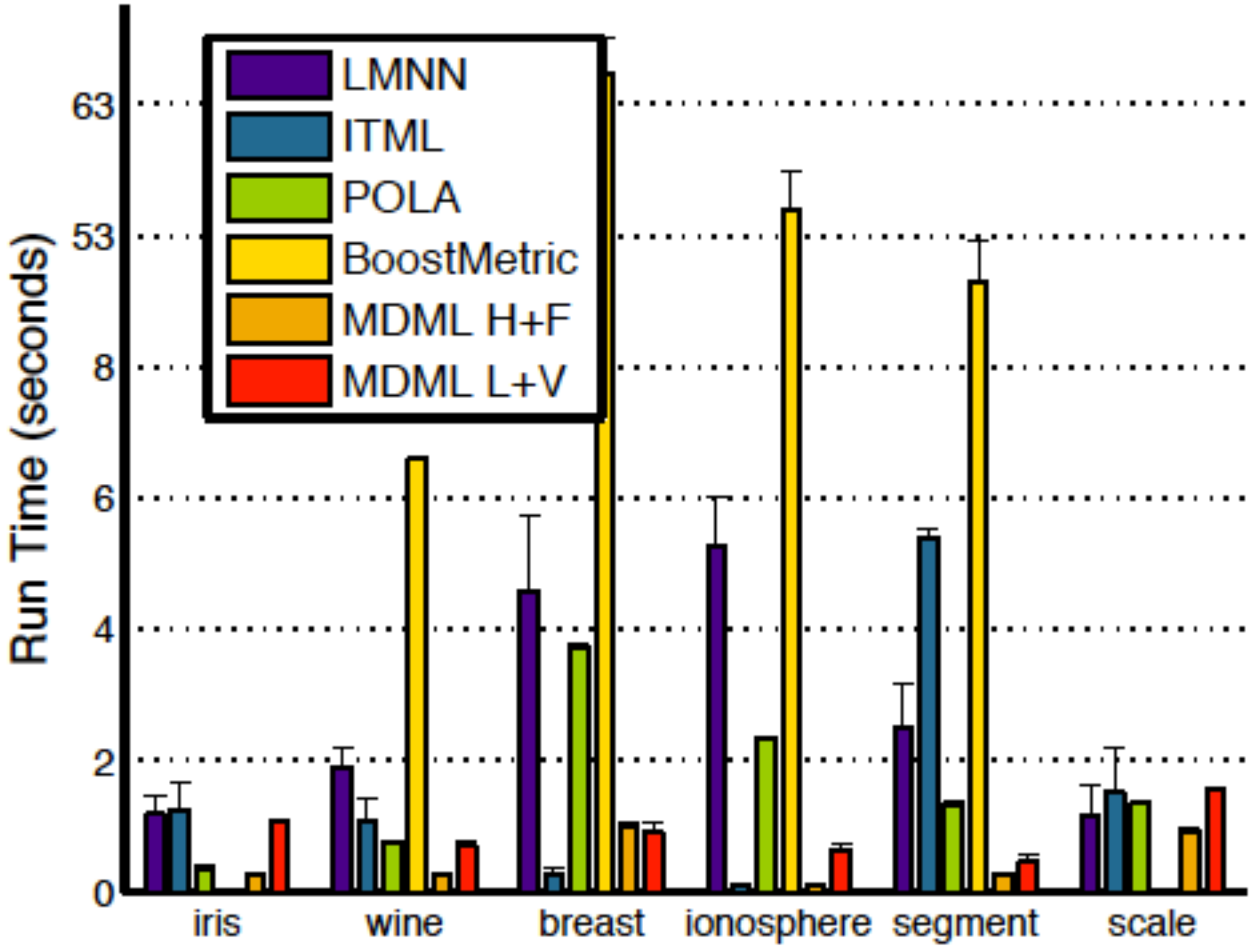
Data set	#train	#test	#dim	#trn pairs	# classes
iris	105	45	4	630	3
wine	123	55	13	738	3
scale	436	189	4	2616	3
segment	147	63	19	882	7
breast	397	172	30	2382	2
ionosphere	245	106	34	1470	2

Test Error on Benchmark Data

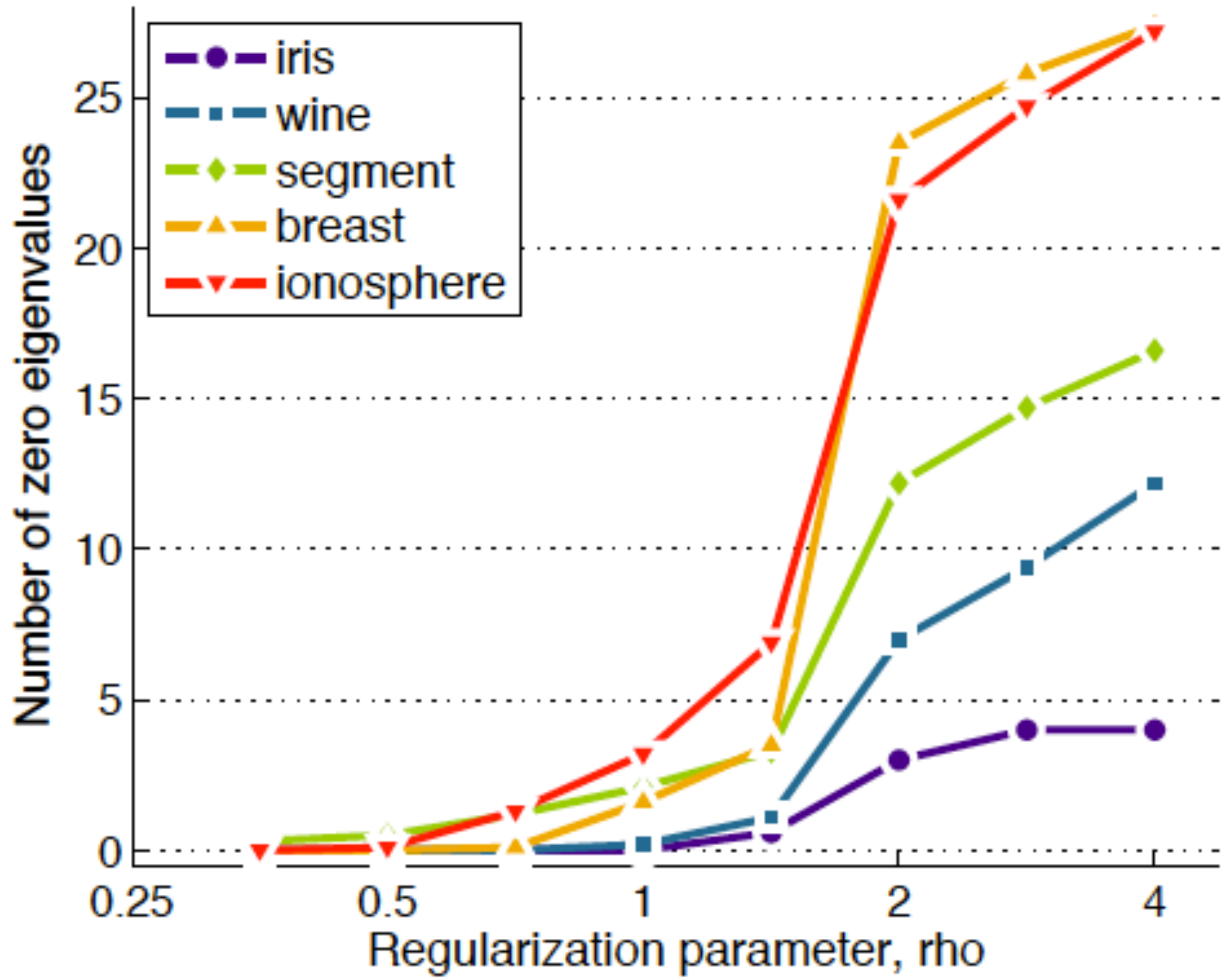


WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Run Times on Benchmark Data

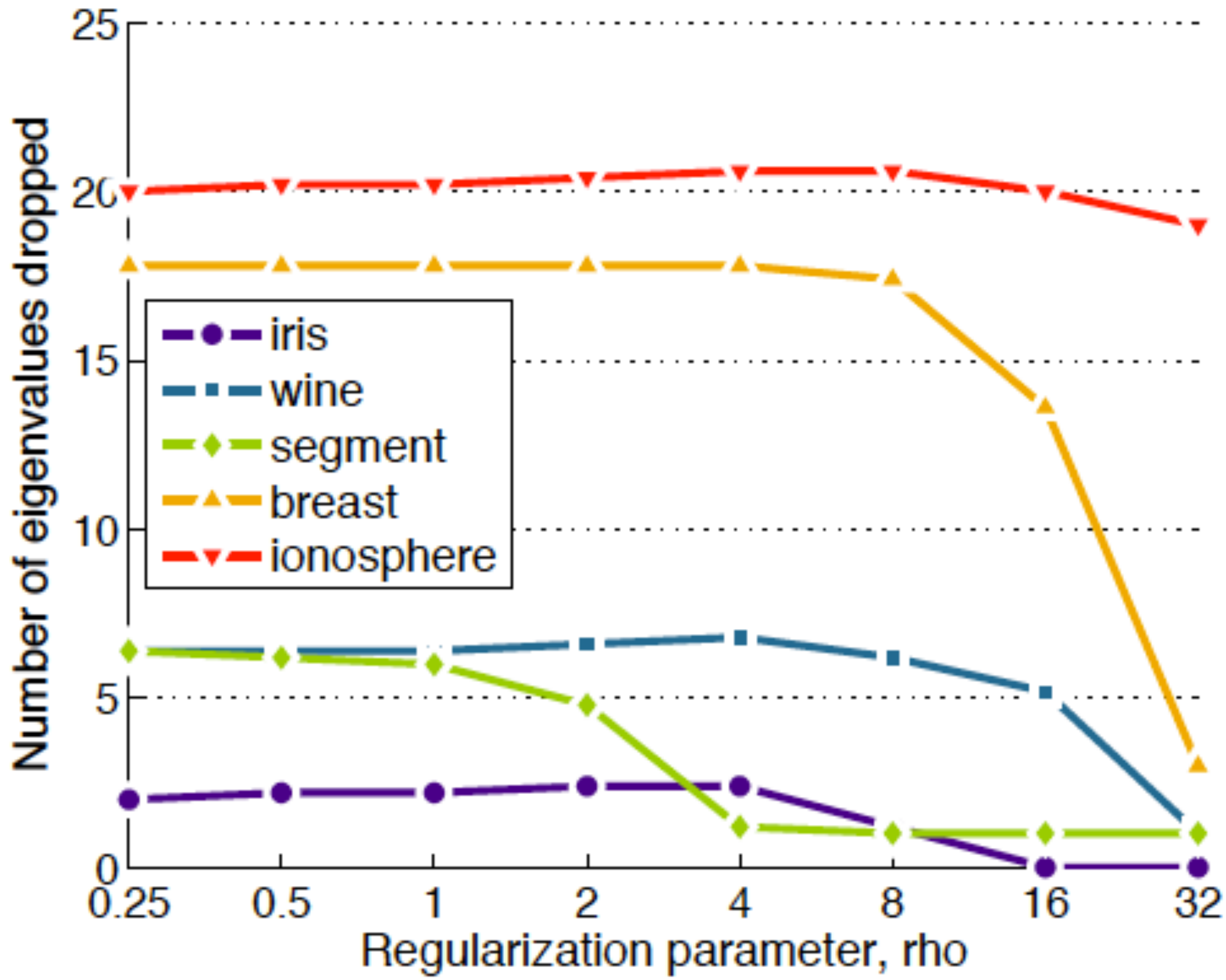


Feature Selection for MDML H+F



WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Feature Selection for MDML L+V



*** eigen-values that account for 90% of the cumulative energy are kept; remaining eigen-values are dropped (similar to PCA)*

OptDigits Data Set

- **Optical Recognition of Handwritten Digits**
  - *64d, 10 classes*
  - *3823 training points and 1797 test points*
  - *11, 469 similar pairs; 11, 469 dissimilar pairs*

Data set	Test Error (%)	Run Time (seconds)	Non-zero features	Num. feats. for 90% energy
LMNN	1.669	54.213	30	20
ITML	5.509	25.745	62	43
POLA	2.282	14.607	53	40
BoostMetric	1.758	2072.427	62	19
MDML H+F	1.892	15.232	26	22
MDML L+V	1.948	13.768	62	29

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Outline

- Introduction
- Mirror Descent for Metric Learning
  – Formulation
  – Loss Functions and Bregman Functions
  – Closed-form Updates
  – Efficient Implementation
- Experiments
  – Results: Benchmark Data Sets
  – Results: OptDigits Data Set
- **Conclusions**

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Conclusions

- **Unifying framework for metric learning**. Different algorithms from various Bregman and loss functions.
- **Scalability**. Updates require rank-1modification of the EVD at each iteration; **implemented efficiently** and **embarrassingly parallel**.
- **Sparse metric**. Minimizing trace norm ensures that M **is sparse** in its eigen-spectrum; only $r < n$ EVs used
- **Kernelizable**.

The authors gratefully acknowledge the support of **Defense Advanced Research Projects Agency** (DARPA) Machine Reading Program under **Air Force Research Laboratory** (AFRL) prime contract no. FA8750-09-C-0181, and the **National Institutes of Health** under the **National Library of Medicine** grant no. NLM R01-LM008796. *The authors would also like to acknowledge anonymous reviewers for their invaluable comments*.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the view of the DARPA, AFRL, or the US government.

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON