Mirror Descent for Metric Learning

Gautam Kunapuli — University of Wisconsin–Madison
Jude W. Shavlik — University of Wisconsin–Madison

The authors gratefully acknowledge DARPA AFRL prime contract FA8750-09-C-0181, and NIH grant NLM R01-LM008796. 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.

Formulating the Problem

We **incrementally learn a pseudo-metric**, $d_M(\mathbf{x}, \mathbf{z})^2 = (\mathbf{x}-\mathbf{z})'M(\mathbf{x}-\mathbf{z})$ given triplets of the form $(\mathbf{x}_t, \mathbf{z}_t, y_t)_{t=1}^T$. The label $y_t = \pm 1$ indicates that $\mathbf{x}_t$ is similar/dissimilar to $\mathbf{z}_t$. where $M \subseteq \mathbb{S}_+^n$. We can introduce the **margin function** [4]:

$$m(\mathbf{x}_t, \mathbf{z}_t, y_t) = y_t \left(\mu - (\mathbf{x}_t - \mathbf{z}_t)'M(\mathbf{x}_t - \mathbf{z}_t) \right),$$

which allows us to define loss for a sample $(\mathbf{x}_t, \mathbf{z}_t, y_t)$; for instance, the hinge loss: $\ell_t(M, \mu) = \max\{0, 1 - m(\mathbf{x}_t, \mathbf{z}_t)\}$. We also add a regularization function $r(M) = |||M|||$, **the trace-norm** of M i.e., the sum of the singular values of M (for some $\rho > 0$) yields sparsity in the singular value spectrum of M, thus minimizing the rank of M:

$$\min_{M \succeq 0, \mu \geq 1} \frac{1}{T} \sum_{t=1}^{T} \ell_t(M, \mu) + r(M),$$

Mirror Descent for Metric Learning

Duchi et al., [2] generalized **mirror descent** to the case where the functions $\phi_t = \ell_t + r$ are composite, consisting of loss and regularization terms. In composite mirror descent (COMID), the ℓ_t is linearized, while r is not. We derive **generalized update rules for a general loss function and Bregman divergence**:

$$M_{t+1} = \arg\min_{M \succeq 0} \; B_\psi(M, M_t) + \eta \langle \nabla_M \ell_t(M_t, \mu_t), M - M_t \rangle + \eta\rho\, |||M|||,$$
$$\mu_{t+1} = \arg\min_{\mu \geq 1} \; B_\psi(\mu, \mu_t) + \eta \nabla_\mu \ell_t(M_t, \mu_t)' (\mu - \mu_t).$$

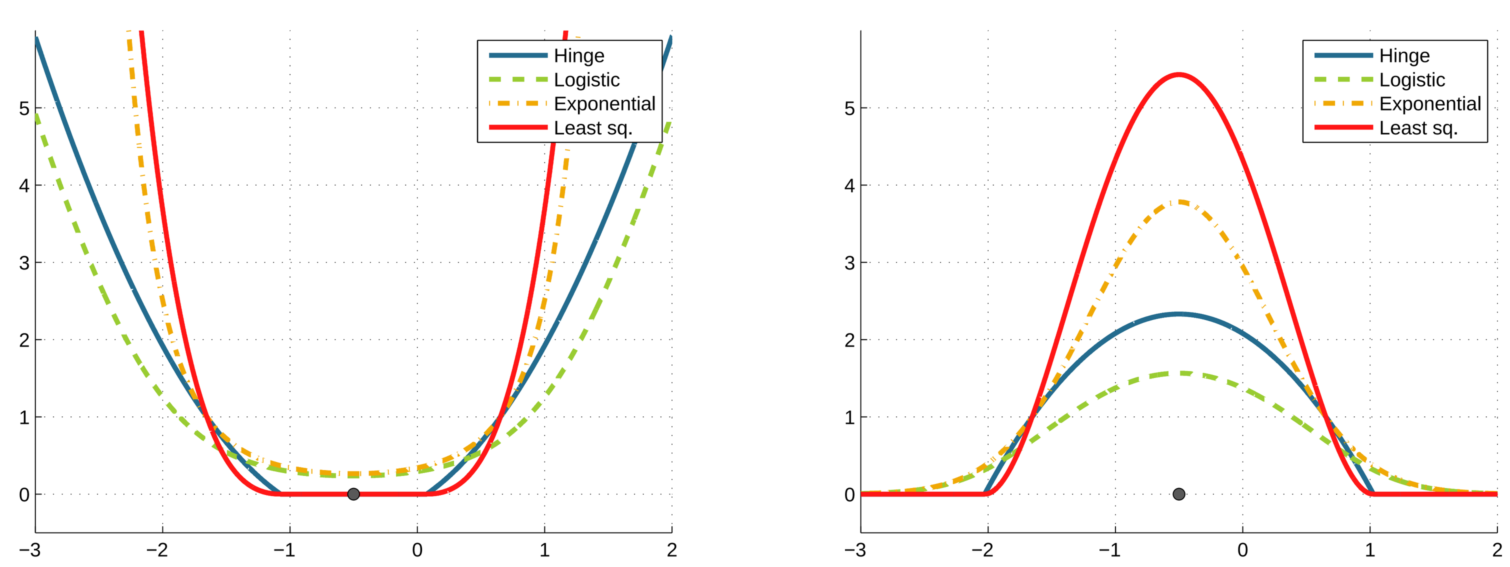
1. **Unifying framework**. Different algorithms arise from various Bregman and loss functions. E.g., using Euclidean distance and relative entropy results in additive and multiplicative updates respectively.
2. **Scalability**. Update rules require rank-one modification of the EVD of $M = V\Lambda V'$; this can be **implemented efficiently** and is **embarrassingly parallel**.
3. **Sparse metric**. The **trace norm** is $|||X||| = \mathbf{e}'|\boldsymbol{\lambda}|$, where $\boldsymbol{\lambda}$ are the EVs of X. Minimizing the trace norm ensures that M is **sparse in its eigenspectrum** i.e., only $r < n$ eigenvalues are used in calculating distances: $\tilde{L} = V_r\sqrt{\Lambda_r}$.
4. **Kernelizable**. The techniques of Chatpatanasiri et al., [1] can be applied here to kernelize it and learn nonlinear metrics.

Bregman Functions and Loss Functions

We consider the following Bregman functions. The squared p-norms $\psi(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|_p^2$ are strongly convex and induce the **squared-Frobenius distance** i.e., $B_\psi(X, Z) = \frac{1}{2}\|X - Z\|_F^2$. The function $\psi(\mathbf{x}) = \sum_i x_i \log x_i - x_i$ induces the **von Neumann divergence**, $B_\psi(X, Y) = \mathrm{tr}\,(X \log X - X \log Y - X + Y)$.

The formulation admits several loss functions. If a loss function is Lipschitz, we obtain algorithms that are characterized by $O(\sqrt{T})$ regret. In the tables below, $\mathbf{u}_t = \mathbf{x}_t - \mathbf{z}_t$.

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	$\log(1 + \exp(-m_t))$	$\frac{\exp(-m_t)}{1+\exp(-m_t)} (y_t \mathbf{u}_t \mathbf{u}_t')$



Different loss functions around $x = -0.5$; **(left)** when $(\mathbf{x}_t, \mathbf{z}_t)$ are similar ($y_t = 1$); **(right)** when $(\mathbf{x}_t, \mathbf{z}_t)$ are dissimilar ($y_t = -1$).

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:

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

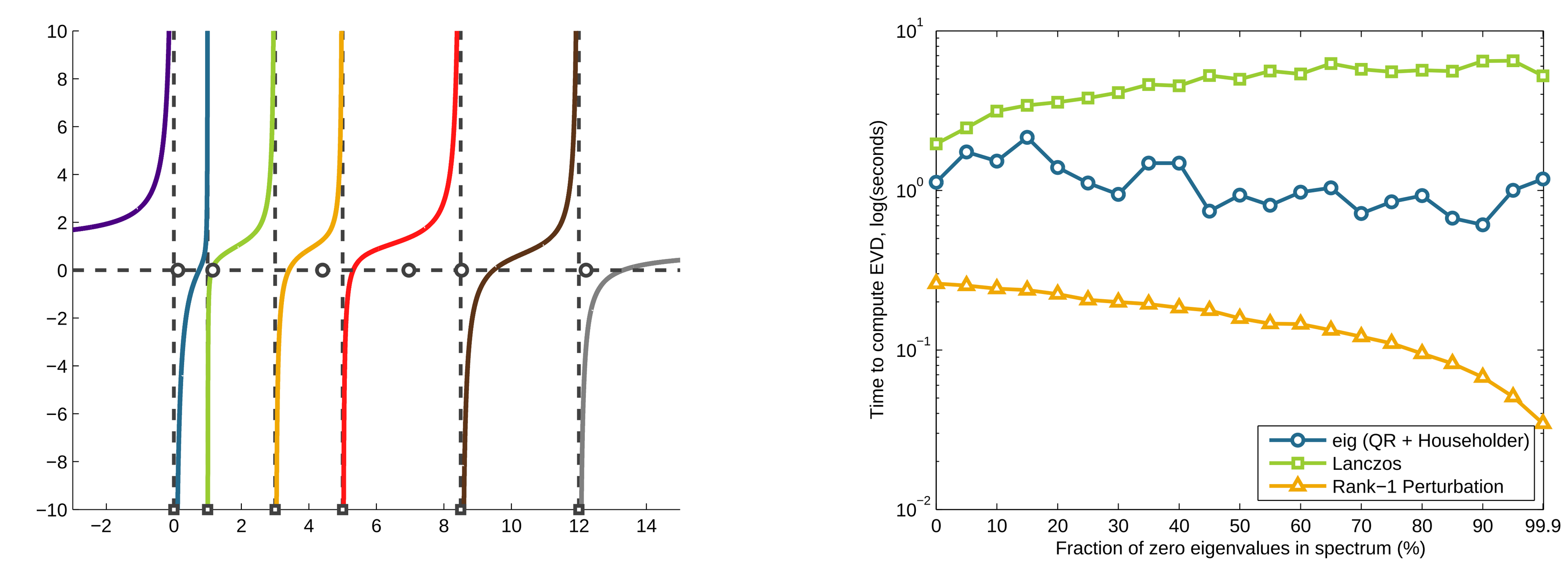
Mirror Descent for Metric Learning

input: data $(\mathbf{x}_t, \mathbf{z}_t, y_t)_{t=1}^T$, parameters ρ, $\eta > 0$
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 $\mu_{t+1} = \max\left(\nabla\psi^{-1}(\nabla\psi(\mu_t) - \eta \nabla \ell_t(M_t, \mu_t)),\ 1 \right)$
end for

Computing EVD Efficiently

We have $M_{t+1} = V_t \nabla\psi(\Lambda_t) V_t' - \alpha \mathbf{u}_t \mathbf{u}_t'$, a *rank-one update* of the EVD at iteration t.

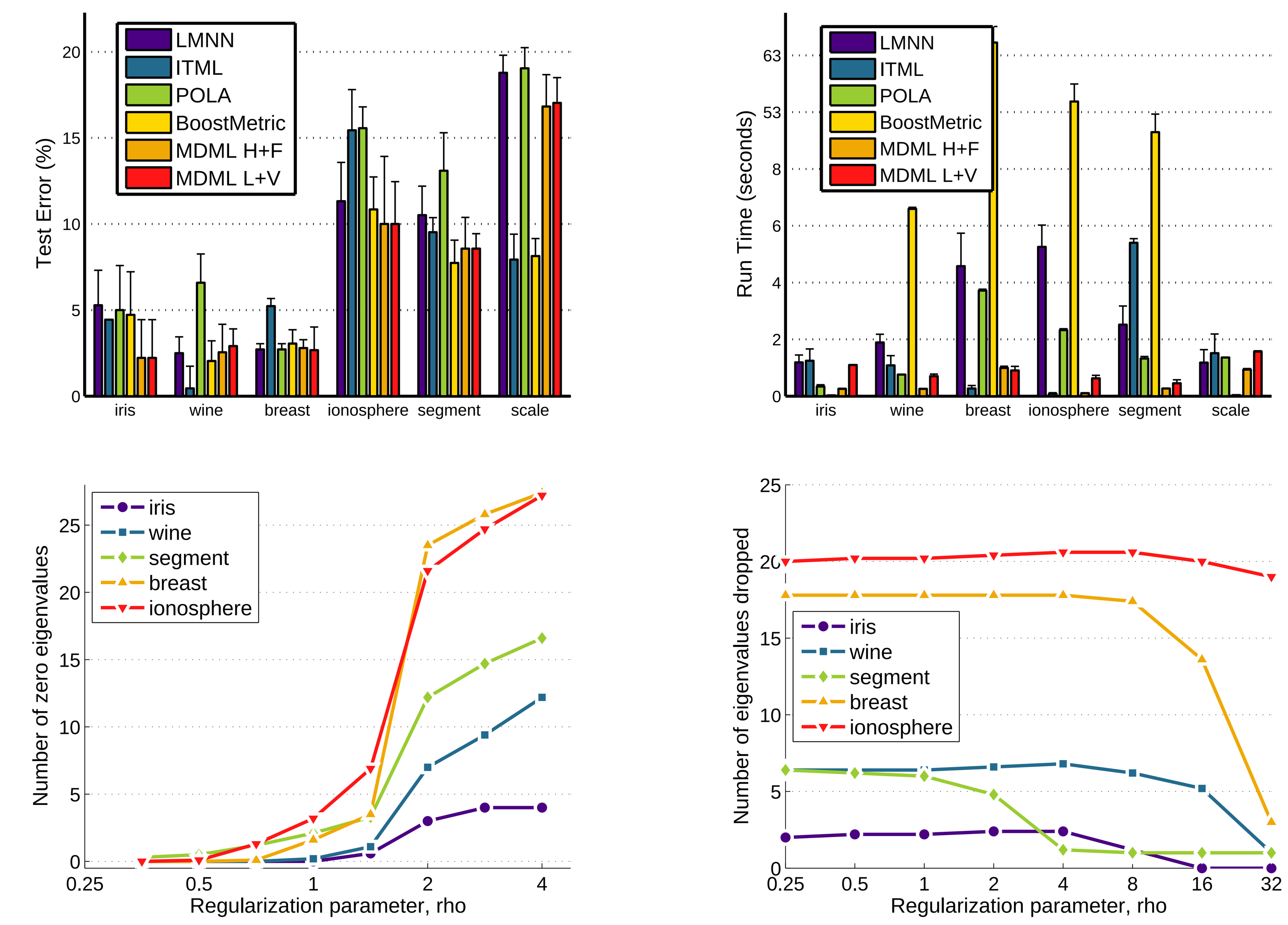
- **Eigenvalue Interlacing**. The EVs of M_t, M_{t+1} interlace; each EV can be computed independently from the secular equation. General root-finding techniques such as Newton may result in non-orthogonal eigenvectors; we adopt the rational interpolation approach of Gu and Eisenstat [3].
- **Learning rate**. An adaptive rate, $\eta_t = \eta/\sqrt{t}$ gives $O(\sqrt{T})$ regret.
- **Low Rank Learning with von Neumann divergence**. This is undefined for low-rank matrices; we update with the *reduced eigendecomposition*, $M_t = \tilde{V}_t \tilde{\Lambda}_t \tilde{V}_t'$. Also, in this case, M_t's smallest EVs are all 1, resulting in full rank; we still perform feature selection by selecting the r largest EVs, similar to PCA.



(left) Interlacing eigenvalues of a matrix and its rank-one perturbation; **(right)** EVD algorithms for randomly generated 500d matrices, over increasing spectrum sparsity.

Experiments: Benchmark Data Sets

We consider two algorithms: an **additive algorithm with hinge loss and Frobenius** (MDML H+F), and a **multiplicative algorithm with logistic loss and von Neumann** (MDML L+V). They are compared to four metric learning approaches: LMNN, ITML, BoostMetric and POLA [4].



References

[1] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul. On kernelization of supervised mahalanobis distance learners. *Computing Research Repoisitory (CoRR)*, abs/0804.1441, 2008.

[2] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, 2010.

[3] Ming Gu and Stanley C. Eisenstat. A stable and efficient algorithm for the rank-one modification of the symmetric eigenproblem. *SIAM Journal on Matrix Analysis and Applications*, 15(4), 1994.

[4] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML'04*, 2004.