MIRROR DESCENT FOR METRIC LEARNING

Gautam Kunapuli
University of Wisconsin–Madison
kunapuli@wisc.edu

Jude W. Shavlik
University of Wisconsin–Madison
shavlik@cs.wisc.edu

The authors gratefully acknowledge support of the National Institute of Health under NLM grant R01-LM008796.

Abstract. We propose a unified approach to Mahalanobis metric learning: an online, regularized, positive semi-definite matrix learning problem, whose update rules can be derived using the composite objective mirror descent (COMID) framework. This approach admits different Bregman and loss functions, which yields several different classes of algorithms. The most novel contribution is the trace norm regularization, which yields a metric sparse in its eigenspectrum, thus performing feature selection. The regularized update rules are parallelizable and can be computed efficiently. The proposed approach is also kernelizable, which allows for metric learning in nonlinear domains.

Formulating the Problem

The goal is to incrementally learn a pseudo-metric, given triplets of the form $(\mathbf{x}_t, \mathbf{z}_t, y_t)_{t=1}^T$. The label $y_t = 1$ indicates that $\mathbf{x}_t$ is similar to $\mathbf{z}_t$ and $y = -1$ indicates dissimilarity:

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

where $L \in \mathbb{R}^{n\times n}$ and $M \subseteq \mathbb{S}^n_+$, the space of symmetric, positive semi-definite matrices. To avoid $O(k^2)$ pairwise constraints from k triplets, the margin $\gamma = 2$ can be fixed and a bias term $\mu \in \mathbb{R}$, is introduced:

$$\left.\begin{array}{l} \forall(\mathbf{x}_1, \mathbf{z}_1, y_1 = +1) \;\Rightarrow\; d_M(\mathbf{x},\, \mathbf{z})^2 \;\le\; \mu - 1, \\ \forall(\mathbf{x}_2, \mathbf{z}_2, y_2 = -1) \;\Rightarrow\; d_M(\mathbf{x},\, \mathbf{z})^2 \;\ge\; \mu + 1. \end{array}\right\} \quad \Rightarrow \quad y_t(\mu - d_M(\mathbf{x}_t,\, \mathbf{z}_t)^2) \;\ge\; 1.$$

We can introduce the **margin function** [4], which allows us to define several loss functions for a sample $(\mathbf{x}_t, \mathbf{z}_t, y_t)$; for instance, the hinge loss:

$$m(\mathbf{x}_t, \mathbf{z}_t, y_t) = y_t(\mu - d_M(\mathbf{x}_t,\, \mathbf{z}_t)^2) = y_t\left(\mu - (\mathbf{x}_t - \mathbf{z}_t)' M (\mathbf{x}_t - \mathbf{z}_t)\right).$$

$$\ell_t(M, \mu) = \max\{\, 0,\, 1 - m(\mathbf{x}_t, \mathbf{z}_t)\,\}.$$

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:

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

Mirror Descent for Metric Learning

The **mirror descent algorithm** [1] is an iterative proximal gradient method for minimizing a convex function, $\phi : \Omega \to \mathbb{R}$. Duchi et al., [3] generalized mirror descent to the case where the functions $\phi_t = \ell_t + r$ are composite, consisting of loss and regularization terms. The subtle difference between the two is that the loss function ℓ_t is linearized, while the regularization r is not.

We derive generalized update rules for a general loss function and Bregman divergence:

$$M_{t+1} = \arg\min_{M\succeq 0} \quad B_\psi(M,\, M_t) + \eta\,\langle\, \nabla_M \ell_t(M_t, \mu_t),\, M - M_t\,\rangle \quad + \eta\,\rho\; |||\, M|||, \tag{2}$$

$$\mu_{t+1} = \arg\min_{\mu\ge 1} \quad B_\psi(\mu,\, \mu_t) + \eta\,\nabla_\mu \ell_t(M_t, \mu_t)'\,(\mu - \mu_t), \tag{3}$$

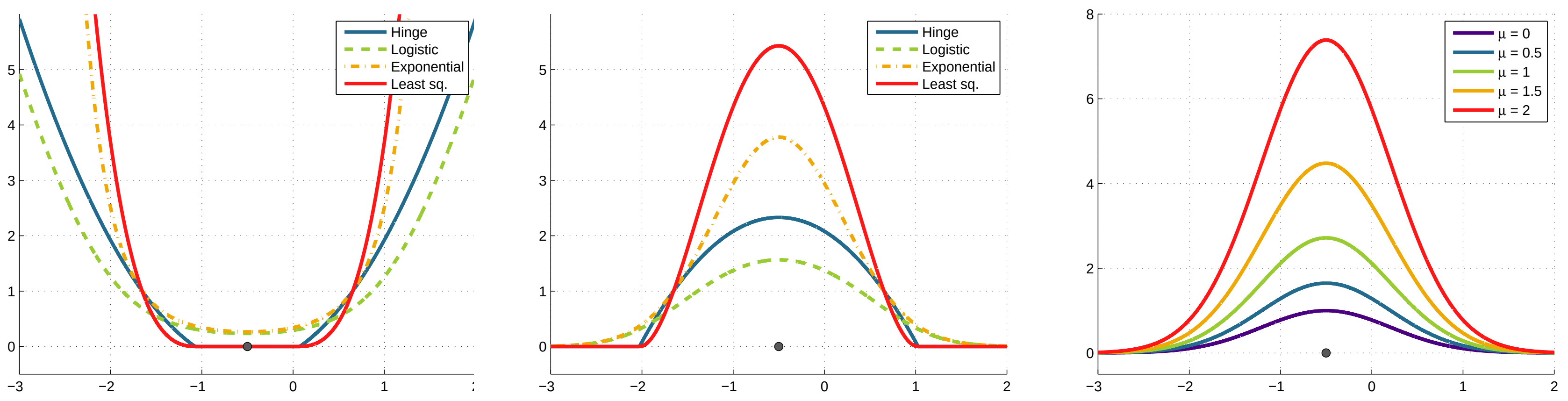
where $\eta > 0$ is the learning rate. **This formulation has several advantages**:

1. **Unifying framework**. Different algorithms arise from different Bregman and loss functions. E.g., using Euclidean distance and relative entropy results in additive and multiplicative updates respectively.
2. **Scalable to large data sets**. The update of the eigendecomposition of $M = V\Lambda V'$, is required at each step. This is **rank-one** and can be implemented very efficiently. The update is also **embarrasingly parallel**.
3. **Trace-norm regularization produces sparse metric**. The trace norm ensures that $M = L'L$ is sparse in its eigenspectrum. If the solution only has $r < n$ eigenvalues, the reduced eigendecomposition can be used in calculating the distances: $\tilde{L} = V_r \Lambda_r$, performing input-space feature selection.
4. **Kernelizable for nonlinear metric learning**. Recently, Chatpatanasiri et al., [2] showed various techniques of kernelizing some popular metric learning approaches. Their results are easily applied to this approach, which can then be applied to learn nonlinear metrics.

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



(**left**) When $(\mathbf{x}_t,\, \mathbf{z}_t)$ are labeled similar ($y_t = 1$): as $\mu \ge 1$ increases, so does the width of the insensitivity of each loss function so that more and more dissimilar points are not penalized; (**center**) When $(\mathbf{x}_t,\, \mathbf{z}_t)$ are labeled dissimilar ($y_t = -1$): a dissimilar pair of points close to each other are heavily penalized; (**right**) μ controls the with of sensitivity around the point of interest; shown for the exponential loss function.

Bregman Functions

The squared Euclidean distance produces an additive update rule; the KL divergence produces a multiplicative update rule; and the Burg divergence produces an inversive update rule.

Divergence	$B_\psi(\mu, \mu_t)$	$\nabla_\mu B_\psi(\mu, \mu_t)$	$B_\psi(M, M_t)$	$\nabla_M B_\psi(M, M_t)$
Euclidean/Frobenius	$\frac{1}{2}(\mu - \mu_t)^2$	$(\mu - \mu_t)$	$\frac{1}{2}\,\lVert M - M_t \rVert_F^2$	$(M - M_t)$
KL/von Neumann	$\mu \log\frac{\mu}{\mu_t} - \mu$	$\log\frac{\mu}{\mu_t}$	**tr** $(M\log M - M\log M_t - M)$	$\log M - \log M_t$
Itakura-Saito/Burg	$\frac{\mu}{\mu_t} - \log\frac{\mu}{\mu_t}$	$\frac{1}{\mu_t} - \frac{1}{\mu}$	**tr** $M M_t^{-1} - \log\det M M_t^{-1}$	$M_t^{-1} - M^{-1}$

Generalized Update Rules

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

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

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

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

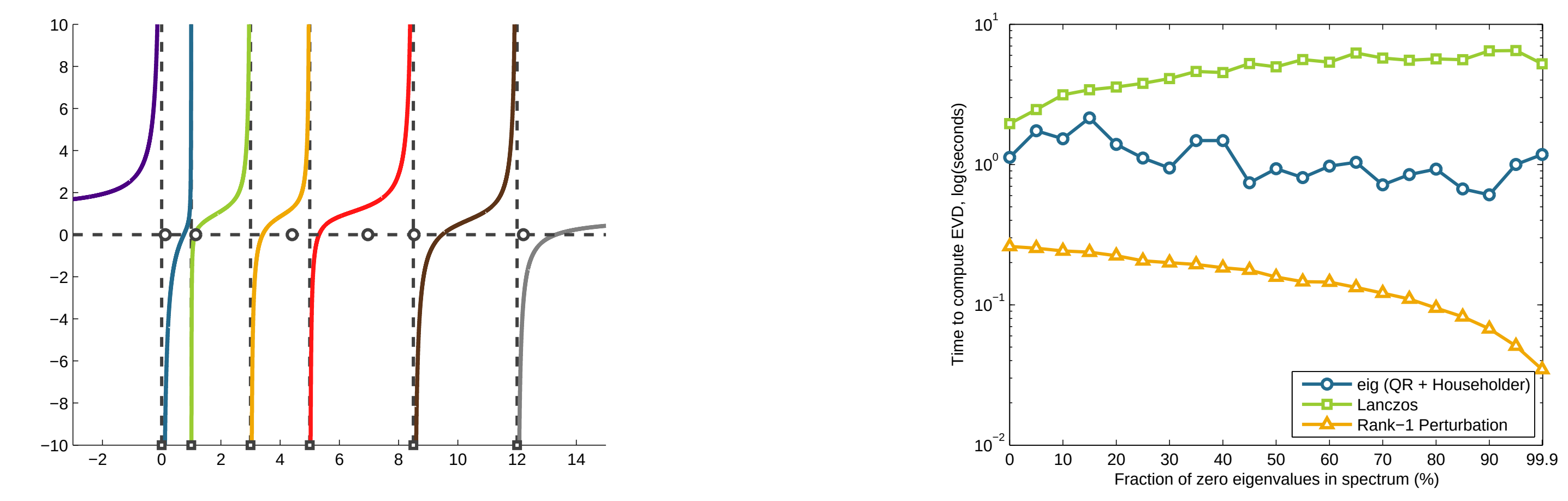
A Generalized Metric Learning Algorithm with Trace-Norm Regularization:

```
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 function $\ell(M, \mu; \mathbf{x},\ \mathbf{z},\ y)$
initialize: $M_0 = I_n\,\mathrm{diag}(\mathbf{0}_n)\,I_n$, $\mu_0 = 0$
for $(\mathbf{x}^t,\ \mathbf{z}_t\, y_t)$ do
    let $\mathbf{u}_t = \mathbf{x}_t - \mathbf{z}_t$
    compute gradients of loss $\nabla_M \ell_t = \alpha_t \mathbf{u}_t \mathbf{u}_t'$ and $\nabla_\mu \ell_t = -\alpha_t$
    write $\nabla\psi(M_t) = V_t \nabla\psi(\Lambda_t) V_t'$
    gradient step $\nabla\psi(M_{t+\frac{1}{2}}) \;=\; V_t \nabla\psi(\Lambda_t) V_t' - \alpha \mathbf{u}_t \mathbf{u}_t'$
    compute evd $\nabla\psi(M_{t+\frac{1}{2}}) \;=\; V_{t+1}\,\Lambda_{t+1}\,V_{t+1}'$
    evd shrinkage $M_{t+1} \;=\; V_{t+1}\,\nabla\psi^{-1}\left(\,S_{\eta\rho}(\Lambda_{t+1})\,\right)\,V_{t+1}'$
    project on to $\mathbb{S}^n_+$ for Frobenius, Burg divergences
    margin update $\mu_{t+1} = \max\left\{\,\nabla\psi^{-1}\left(\nabla\psi(\mu_t) - \eta\,\nabla_\mu \ell_t\right),\ 1\right\}$
end for
```

Computing the Eigen-Decomposition Efficiently

Steps 8 and 9 above are implemented together, where a symmetric rank-one update $M = V\mathrm{diag}(\boldsymbol{\lambda})V' + \alpha\mathbf{u}\mathbf{u} = W\mathrm{diag}(\boldsymbol{\mu})W'$ is computed. Assume $\lambda_1 = \cdots = \lambda_k = \bar{\lambda} < \lambda_{k+1} < \ldots < \lambda_n$ and $\alpha > 0$. By the **eigenvalue interleaving theorem**, we have $\{\mu_i = \lambda_i\}_{i=1}^{k-1}$. We only have to update $n - k + 1$ eigenvalues.

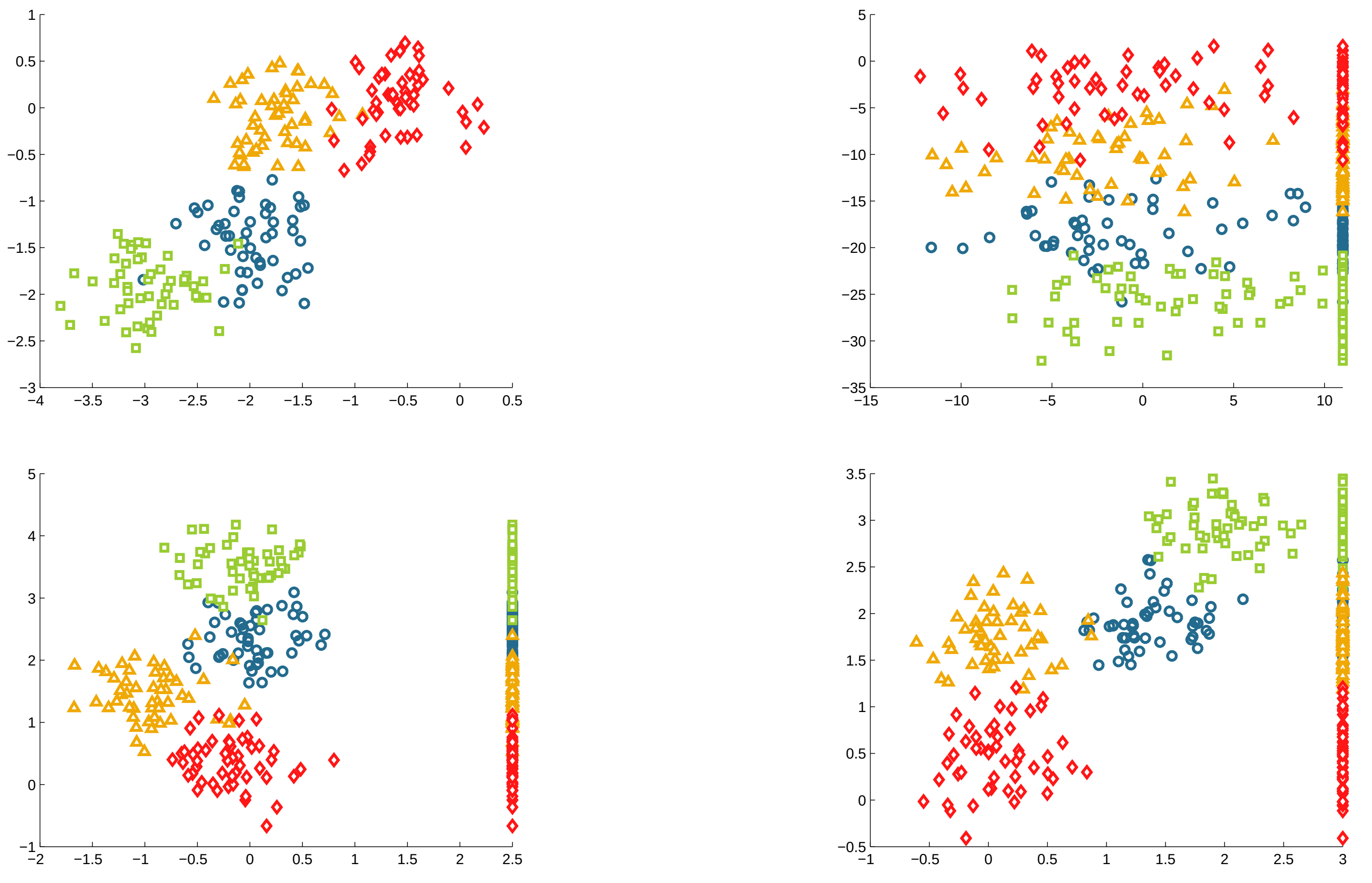
```
input: $\mathbf{t} = V'\mathbf{u}$,
pre-process: compute $\tilde{\mathbf{t}} = \left[\tilde{\mathbf{t}}_{1:k}',\ \mathbf{t}_{k+1:n}'\right]'$, and $\tilde{V} = \left[\,\tilde{V}_{1:k}\ V_{k+1:n}\,\right]$ where

    $\tilde{\mathbf{t}}_{1:k} = \left(I_k - 2\frac{\mathbf{s}\mathbf{s}'}{\mathbf{s}'\mathbf{s}}\right)\mathbf{t}_{1:k}, \quad \tilde{V}_{1:k} = V_{1:k}\left(I_k - 2\frac{\mathbf{s}\mathbf{s}'}{\mathbf{s}'\mathbf{s}}\right),$

    and $\mathbf{s} = \mathbf{t}_{1:k} + \|\mathbf{t}_{1:k}\|_2\,\mathbf{e}_1$ is a Householder reflector such that $\tilde{\mathbf{t}}_{1:k} = [-\|\mathbf{t}_{1:k}\|_2,\ 0\ \ldots,\ 0]'$. This does not
    change the problem since $M = \tilde{V}\mathrm{diag}(\boldsymbol{\lambda})\tilde{V}'$ and $\mathbf{u} = \tilde{V}'\tilde{\mathbf{t}}$
compute eigenvalues: solve the secular equation for $\mu_i$ between $[\lambda_i, \lambda_{i+1}]$, for $i = k, \ldots, n$

    $f(\mu) \;:=\; 1 - \alpha\tilde{\mathbf{t}}'(\mu I_n - \mathrm{diag}(\boldsymbol{\lambda}))^{-1}\tilde{\mathbf{t}} \;=\; 0,$

compute eigenvectors: $\mathbf{w}_i = V(\mu_i I_n - \mathrm{diag}(\boldsymbol{\lambda}))^{-1}\tilde{\mathbf{t}}$ and normalize
```



(**left**) The **secular equation, with interleaving eigenvalues** λ_i of a matrix $M \in \mathbb{S}^6$ with the eigenvalues μ_i of a rank-one perturbation $\tilde{M} = M + \alpha\mathbf{u}\mathbf{u}'$; (**right**) Comparing eigenvalue algorithms: EVDs are computed for 10 random matrices in $\mathbb{S}^{500}$, over increasing sparsity in the spectrum. The **rank-one perturbation approach outperforms standard EVD approaches** as it exploits structure and eigenvalue interleaving. Also, its **computational cost sharply drops with increasing sparsity**.

A Simple Example

A 4-class **2d data set; 8 spurious dimensions are added** to create a 10d training set (**top left**); Data projected on to the top two eigenvectors of the learned M, and on the far right, data projected onto the largest eigenvector of the learned M, for (**top right**) Frobenius + hinge loss; (**bottom left**) von Neumann + logistic loss; (**bottom right**) Burg + logistic loss.



References

[1] A. Beck and M. Teboulle. Mirror descent nonlinear projected subgrad. methods for convex optimization. *Oper. Rsch. Letters*, 31:167–175, 2003.

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

[3] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, pages 14–26, 2010.

[4] S. Shalev-Shwartz, Y. Singer, and A. Y. Ng. Online and batch learning of pseudo-metrics. In *ICML'04*, pages 94–102, 2004.