Inductive Regularized Learning of Kernel Functions: Supplementary Material

Prateek Jain
Microsoft Research Bangalore
Bangalore, India
prajain@microsoft.com

Brian Kulis
UC Berkeley EECS and ICSI
Berkeley, CA, USA
kulis@eecs.berkeley.edu

Inderjit Dhillon
UT Austin Dept. of Computer Sciences
Austin, TX, USA
inderjit@cs.utexas.edu

Appendix A

In this section we provide detailed proofs for Theorems 1-3.

Recall that our kernel matrix learning problem is given by

$$\min_{K_W \succeq 0} f(K^{-1/2} K_W K^{-1/2}) \qquad \text{s.t. } g_i(K_W) \le b_i,\ 1 \le i \le m, \tag{1}$$

while our linear transformation kernel learning problem is given by

$$\min_{W \succeq 0} f(W) \qquad \text{s.t. } g_i(\Phi^T W \Phi) \le b_i,\ 1 \le i \le m. \tag{2}$$

First we introduce and analyze an auxiliary optimization problem that will help in proving the main theorems. Consider the following problem:

$$\begin{aligned} \min_{W \succeq 0, L} \quad & f(W) \\ \text{s.t.} \quad & g_i(\Phi^T W \Phi) \le b_i,\ 1 \le i \le m, \\ & W = \alpha I^d + U L U^T, \end{aligned} \tag{3}$$

where $L \in \mathbb{R}^{k \times k}$, $U \in \mathbb{R}^{d \times k}$ is an orthogonal matrix, and I^d is the $d \times d$ identity matrix. In general, k can be significantly smaller than $\min(n, d)$. Note that the above problem is identical to (2) except for an added constraint $W = \alpha I^d + U L U^T$. We now show that (3) is equivalent to a problem over $k \times k$ matrices. In particular, (3) is equivalent to (4) defined below.

Lemma 1. *Let f be a spectral function (see Defintion 3.1) and let α be the global minima for the corresponding scalar function f_s. Then,* (3) *is equivalent to:*

$$\begin{aligned} \min_{L} \quad & f(\alpha I^k + L), \\ \textit{s.t.} \quad & g_i(\alpha \Phi^T \Phi + \Phi^T U L U^T \Phi) \le b_i,\ 1 \le i \le m, \\ & L \succeq -\alpha I^k. \end{aligned} \tag{4}$$

Proof. The last constraint in (3) asserts that $W = \alpha I^d + U L U^T$, which implies that there is a one-to-one mapping between W and L: given W, L can be computed and vice-versa. As a result, we can eliminate the variable W from (3) by substituting $\alpha I^d + ULU^T$ for W (via the last constraint in (3)). The resulting optimization problem is:

$$\begin{aligned}
\min_{L} \quad & f(\alpha I + ULU^T), \\
\text{s.t.} \quad & g_i(\alpha \Phi^T \Phi + \Phi^T ULU^T \Phi) \leq b_i, \ 1 \leq i \leq m, \\
& L \succeq -\alpha I^k.
\end{aligned} \tag{5}$$

Note that (4) and (5) are the same except for their objective functions. Below, we show that both the objective functions are equal up to a constant, so they are interchangable in the optimization problem. Let $U' \in \mathbb{R}^{d \times d}$ be an orthonormal matrix obtained by completing the basis represented by U, i.e., $U' = [U \ U_\perp]$ for some $U_\perp \in \mathbb{R}^{d \times (d-k)}$ s.t. $U^T U_\perp = 0$ and $U_\perp^T U_\perp = I^{d-k}$. Now,

$$W = \alpha I + ULU^T = U' \left(\alpha I + \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} \right) U'^T. \tag{6}$$

It is straightforward to see that for a spectral function f,

$$f(VWV^T) = f(W), \tag{7}$$

where V is an orthogonal matrix. Also, $\forall A, B \in \mathbb{R}^{d \times d}$,

$$f \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} = f(A) + f(B). \tag{8}$$

Using (6), (7), and (8), we get:

$$\begin{aligned}
f(W) = f(\alpha I + ULU^T) &= \left(\alpha U'^T I U' + \begin{bmatrix} L & 0 \\ 0 & 0 \end{bmatrix} \right), \\
&= f\left(\begin{bmatrix} \alpha I + L & 0 \\ 0 & \alpha I \end{bmatrix} \right), \\
&= f(\alpha I + L)) + (d - n) f(\alpha),
\end{aligned} \tag{9}$$

Therefore, the objective functions of (4) and (5) differ by only a constant, i.e., they are equivalent w.r.t. the optimization problem. The lemma follows. $\square$

We now show that for the convex spectral functions (see Definition 3.1) the optimal solution W^* to (2) is of the form $W^* = I + \Phi S \Phi^T$, for some S.

Lemma 2. *Suppose f satisfies the conditions given in Theorem 1. Furthermore, denote the global minima of the corresponding scalar function f_s as α. Then, the optimal solution to* (2) *is of the form $W^* = \alpha I + \Phi S \Phi^T$, where S is a $n \times n$ matrix.*

Proof. Let $W = U \Lambda U^T = \sum_j \lambda_j \boldsymbol{u}_j \boldsymbol{u}_j^T$ be the eigenvalue decomposition of W. Consider a constraint $g_i(\Phi^T W \Phi) \leq b_i$ as specified in (2). Note that if the j-th eigenvector $\boldsymbol{u}_j$ of W is orthogonal to the range space of Φ, i.e. $\Phi^T \boldsymbol{u}_j = 0$, then the corresponding eigenvalue λ_j is not constrained (except for the non-negativity constraint imposed by the positive semi-definiteness constraint). Since the range space of Φ is at most n-dimensional, without loss of generality we can assume that $\lambda_j \geq 0, \forall j > n$ are not constrained by the linear inequality constraints in (2).

Since f satisfies the conditions of Theorem 1, $f(W) = \sum_j f_s(\lambda_j)$. Also, $f_s(\alpha) = \min_x f_s(x)$. Hence, to minimize $f(W)$, we can select $\lambda_j^* = \alpha \geq 0, \forall j > n$ (note that the non-negativity constraint is satisfied for this choice of λ_j). Furthermore, the eigenvectors $\boldsymbol{u}_j$, $\forall j \leq n$, lie in the range space of X, i.e., $\forall j \leq n$, $\boldsymbol{u}_j = X \boldsymbol{z}_j$ for some $\boldsymbol{z}_j \in \mathbb{R}^n$. Therefore,

$$\begin{aligned}
W^* &= \sum_{j=1}^{n} \lambda_j^* \boldsymbol{u}_j^* \boldsymbol{u}_j^{*T} + \alpha \sum_{j=n+1}^{d} \boldsymbol{u}_j^* \boldsymbol{u}_j^{*T}, \\
&= \sum_{j=1}^{n} (\lambda_i^* - \alpha) \boldsymbol{u}_j^* \boldsymbol{u}_j^{*T} + \alpha \sum_{j=1}^{d} \boldsymbol{u}_j^* \boldsymbol{u}_j^{*T}, \\
&= \Phi S^* \Phi^T + \alpha I,
\end{aligned}$$

where $S^* = \sum_{j=1}^{n} (\lambda_j^* - \alpha) \boldsymbol{z}_j^* \boldsymbol{z}_j^{*T}$. $\square$

Now we use Lemmas 1 and 2 to prove Theorem 1.

Proof of Theorem 1. Let $\Phi = U_\Phi \Sigma V_\Phi^T$ be the singular value decomposition (SVD) of Φ. Note that

$$K = \Phi^T \Phi = V_\Phi \Sigma^2 V_\Phi^T.$$

Also, assuming $\Phi \in \mathbb{R}^{d \times n}$ to be full-rank and $d > n$, $V_\Phi V_\Phi^T = I$.

Using Lemma 2, the optimal solution to (2) is restricted to be of the form $W = \alpha I + \Phi S \Phi^T = \alpha I + U_\Phi \Sigma V_\Phi^T S V_\Phi \Sigma U_\Phi^T = \alpha I + U_\Phi V_\Phi^T K^{1/2} S K^{1/2} V_\Phi U_\Phi^T = \alpha I + U_\Phi V_\Phi^T L V_\Phi U_\Phi^T$, where $L = K^{1/2} S K^{1/2}$. Hence, for spectral functions f, (2) is equivalent to (3), so using Lemma 1, (2) is equivalent to (4) with $U = U_\Phi V_\Phi^T$ and $L = K^{1/2} S K^{1/2}$. Also, note that the constraints in (4) can be simplified to:

$$g_i(\alpha \Phi^T \Phi + \Phi^T U L U^T \Phi) \le b_i \equiv g_i(\alpha K + K^{1/2} L K^{1/2}) \le b_i.$$

Now, let $K_W = \alpha K + K^{1/2} L K^{1/2} = \alpha K + KSK$, i.e., $L = K^{-1/2}(K_W - \alpha K) K^{-1/2}$. Theorem 1 now follows by substituting for L in (4). □

Next, we prove Theorem 2.

Proof of Theorem 2. Let $U = K^{1/2} J (J^T K J)^{-1/2}$ and let J be a full rank matrix, then U is an orthgonal matrix. Using (9) we get,

$$f(\alpha I + U (J^T K J)^{1/2} L (J^T K J)^{1/2} U^T) = f(\alpha I + (J^T K J)^{1/2} L (J^T K J)^{1/2}).$$

Now consider a linear constraint specified in (6) (*from main text*), $\mathrm{Tr}(C_i(\alpha K + KJLJ^T K)) \le b_i$. This can be easily simplified to:

$$\mathrm{Tr}(L J^T K C_i K J) \le b_i - \mathrm{Tr}(\alpha K C_i).$$

Similar simple algebraic manipulations to the PSD constraint completes the proof. □

Finally, we prove Theorem 3.

Proof of Theorem 3. Consider the last constraint in (7) (*from main text*):

$$W = \alpha I + \Phi J L J \Phi^T.$$

Let $\Phi = U \Sigma V^T$ be the SVD of Φ. Hence, $W = \alpha I + U V^T V \Sigma V^T J L J V \Sigma V^T V U^T = \alpha I + U V^T K^{1/2} J L J K^{1/2} V U^T$. For disambiguity, rename L as L' and U as U'. Now, clearly (7) (*from main text*) is same as (3) with $U = U' V^T$ and $L = K^{1/2} J L' J K^{1/2}$. Theorem 3 now follows by using Lemma 1 with $L = K^{1/2} J L' J K^{1/2}$. □

Appendix B: Trace-SSIKDR

To recap, the updates for solving (11) (*from main text*) using Uzawa's algorithm are given by:

$$U \Sigma U^T \leftarrow K^{1/2} C K^{1/2}, \tag{10}$$

$$\tilde{K}^t \leftarrow U \max(\Sigma - \tau I, 0) U^T, \tag{11}$$

$$z_i^t \leftarrow z_i^{t-1} - \delta \max(\mathrm{Tr}(C_i K^{1/2} \tilde{K}^t K^{1/2}) - b_i, 0), \forall i, \tag{12}$$

where $C = \sum_i z_i^{t-1} C_i$. We first prove a technical lemma to relate eigenvectors vectors U of matrix $K^{1/2} C K^{1/2}$ and V of the matrix CK.

Lemma 3. *Let $K^{1/2} C K^{1/2} = U_k \Sigma_k U_k^T$, where U_k contains the top-k eigenvectors of $K^{1/2} C K^{1/2}$ and Σ_k contains the top-k eigenvalues of $K^{1/2} C K^{1/2}$. Similarly, let $CK = V_k \Lambda_k V_k^{-1}$, where V_k contains the top-k right eigenvectors of CK and Λ_k contains the top-k eigenvalues of CK. Then,*

$$U_k = K^{1/2} V_k D_k,$$
$$\Sigma_k = \Lambda_k.$$

Note that eigenvalue decomposition is unique up to sign, so we assume that the sign has been set correctly.

Proof. Let $\boldsymbol{v}_i$ be i-th eigenvector of CK. Then, $CK\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$. Multiplying both sides with $K^{1/2}$, we get $K^{1/2}CK^{1/2}K^{1/2}\boldsymbol{v}_i = K^{1/2}\boldsymbol{v}_i$. After normalization we get:

$$(K^{1/2}CK^{1/2})\frac{K^{1/2}\boldsymbol{v}_i}{\boldsymbol{v}_i^T K \boldsymbol{v}_i} = \lambda_i \frac{K^{1/2}\boldsymbol{v}_i}{\boldsymbol{v}_i^T K \boldsymbol{v}_i}$$

Hence, $\frac{K^{1/2}\boldsymbol{v}_i}{\boldsymbol{v}_i^T K \boldsymbol{v}_i} = K^{1/2}\boldsymbol{v}_i / D(i,i)$ is the i-th eigenvector $\boldsymbol{u}_i$ of $K^{1/2}CK^{1/2}$. Also, $\sigma_i = \lambda_i$. $\square$

Using the above lemma and (11), we get

$$\tilde{K} = K^{1/2} V_k D_k \lambda D_k V_k^{-1} K^{1/2}.$$

Therefore, the update for the z variables (see (12)) reduces to:

$$z_i^t \leftarrow z_i^{t-1} - \delta \max(\mathrm{Tr}(C_i K V_k D_k \lambda D_k V_k^{-1} K) - b_i, 0), \forall i.$$

This proves that step 6 of Algorithm 1 is correct, so we do not need to compute the full eigenvalue decompsotion or square-root of the kernel matrix K.