A Convex Feature Learning Formulation for Latent Task Structure Discovery
(Supplementary Material)

Pratik Jawanpuria pratik.j@cse.iitb.ac.in
J. Saketha Nath saketh@cse.iitb.ac.in

Dept. of CSE, IIT Bombay, Mumbai, India

We follow the notation described in the submission. This text may refer to equations/theorems/lemmas in the original submission using the appropriate numbers therein.

A Derivation of partial dual in Theorem 1

We first present a detailed derivation of the partial dual (4). We begin by writing the primal formulation with the proposed regularizer (3).

$$\min_{\mathbf{h},\mathbf{b}} \quad \frac{1}{2}\left(\sum_{v\in\mathcal{V}} d_v \left(\sum_{w\in D(v)} \left(\sum_j \left(\mu\|h_{w0}^j\|_2^2 + \sum_{t\in w}\|h_{tw}^j\|_2^2\right)^{\frac{p}{2}}\right)^{\frac{1}{p}}\right)^{\frac{1}{q}}\right)^2 + C\sum_{t,i}\ell(F_t(\mathbf{x}_{ti}), y_{ti}) \tag{A.1}$$

Without loss of generality, we assume that tasks within each group/node w are arranged in the order of their number (hence t_i before t_j if $t_i < t_j$). Next, we formulate the feature map of an input instance $\mathbf{x}$ of task t (similar to Evgeniou & Pontil (2004)), but for each group w and for each kernel space k^j. For a given input instance $\mathbf{x}$ of task t and feature space ϕ^j, its feature map within a group w (containing task t) will be: $\Phi_w^j(\mathbf{x},t) = (\frac{\phi^j(\mathbf{x})}{\sqrt{\mu}}, \underbrace{\mathbf{0},\ldots,\mathbf{0}}_{\text{tasks before t}}, \phi^j(\mathbf{x}), \underbrace{\mathbf{0},\ldots,\mathbf{0}}_{\text{tasks after t}})$, where $\mathbf{0}$ represents a zero vector in the induced feature space of $\phi^j(\cdot)$. The corresponding model of the group in j^{th} feature space will be: $f_w^j = (\sqrt{\mu}h_{0w}^j, h_{t_1w}^j, \ldots, h_{t_{|w|}w}^j)$. We also define the overall feature map and model of a group w as $\Phi_w = (\Phi_w^1,\ldots,\Phi_w^n)$ and $f_w = (f_w^1,\ldots,f_w^n)$ respectively. From such constructions, it is evident that the the mixed-norm base regularizer $\|\Theta_w\|_p$ within each group w can be written as: $\|\Theta_w\|_p = \left(\sum_j \|f_w^j\|_2^p\right)^{\frac{1}{p}} = \|f_w\|_{p,2}$. Using the above defined feature map, the decision function becomes $F_t(\mathbf{x}) = \sum_{w:t\in T_w}\langle f_w, \Phi_w(\mathbf{x},t)\rangle - b_t$ and formulation (A.1) with hinge loss function can be equivalently rewritten as:

$$\begin{aligned} \min_{\mathbf{f},\xi,\mathbf{b}} \quad & \tfrac{1}{2}\Omega_S(\mathbf{f})^2 + C\sum_{t=1}^T\sum_{i=1}^m \xi_{ti} \\ \text{s.t.} \quad & y_{ti}F_t(\mathbf{x}_{ti}) \geq 1-\xi_{ti},\ \xi_{ti}\geq 0\ \forall\ i,t \end{aligned} \tag{A.2}$$

where $\Omega_S(\mathbf{f}) = \sum_{v\in\mathcal{V}} d_v\|f_{D(v)}\|_{q,p}$ and $\|f_{D(v)}\|_{q,p} = \left(\sum_{w\in D(v)}\|f_w\|_{p,2}^q\right)^{\frac{1}{q}}$.
Note that (A.2) aims at finding an optimal vector in an RKHS. In order to facilitate the application of a suitable representer theorem (Schölkopf & Smola, 2002) for writing down the dual, we employ a variational formulation of $\Omega_S(\mathbf{f})$. We note the following lemma (Micchelli & Pontil, 2005) which will be useful in the derivations:

Lemma A.1. *Let $a_i \geq 0, i = 1, \ldots, d$ and $1 \leq r < \infty$. Let $\Delta_{d,r} = \left\{\theta \in \mathbb{R}^d \mid \theta \geq 0, \sum_{i=1}^d \theta_i^r = 1\right\}$. Then,*

$$\min_{\gamma \in \Delta_{d,r}} \sum_i \frac{a_i}{\gamma_i} = \left(\sum_{i=1}^d a_i^{\frac{r}{r+1}}\right)^{1+\frac{1}{r}} \tag{A.3}$$

and the minimum is attained at

$$\gamma_i = \frac{a_i^{\frac{1}{r+1}}}{\left(\sum_{i=1}^d a_i^{\frac{r}{r+1}}\right)^{\frac{1}{r}}}$$

Here, by convention, $a/0$ is 0 if $a = 0$ and is ∞ otherwise. In the limit $r \to \infty$ the following holds

$$\min_{\gamma \in \Delta_{d,\infty}} \sum_i \frac{a_i}{\gamma_i} = \sum_{i=1}^d a_i \tag{A.4}$$

and equality is attained when $\gamma_i = 1 \; \forall \; a_i > 0$.

Applying the lemma we get the following equalities:

$$\Omega(\mathbf{f})^2 = \left(\sum_{v \in V} d_v \|f_{D(v)}\|_{q,p}\right)^2 = \min_{\eta \in \Delta_{|V|,1}} \sum_{v \in V} \frac{d_v^2 \|f_{D(v)}\|_{q,p}^2}{\eta_v}$$

$$\|\mathbf{f}_{D(v)}\|_{q,p}^2 = \min_{\zeta_v \in \Delta_{|D(v)|,\hat{q}}} \sum_{w \in D(v)} \frac{\|f_w\|_{p,2}^2}{\zeta_{vw}}, \quad \text{where } \hat{q} = \frac{q}{2-q}$$

$$\|f_w\|_{p,2}^2 = \min_{\sigma_w \in \Delta_{n,\hat{p}}} \sum_{j=1}^n \frac{\|f_w^j\|_2^2}{\sigma_w^j}, \quad \text{where } \hat{p} = \frac{p}{2-p}$$

Thus we have the following variational formulation:

$$\Omega(\mathbf{f})^2 = \min_{\eta \in \Delta_{|\mathcal{V}|,1}} \quad \min_{\zeta_v \in \Delta_{|D(v)|,\hat{q}} \forall v \in V} \quad \min_{\sigma_v \in \Delta_{k,\hat{p}} \forall v \in V} \sum_{w \in \mathcal{V}} \tau_w^{-1}(\eta, \zeta) \frac{\|f_w\|_2^2}{\sigma_w^j}$$

where $\tau_w^{-1}(\eta, \zeta) = \sum_{v \in A(w)} \frac{d_v^2}{\eta_v \zeta_{vw}}$. Next, by applying the representer theorem (see also Rakotomamonjy et al. (2008)) and using the notion of dual norm Boyd & Vandenberghe (2004), we derive a partial-dual (dual wrt. variables $\mathbf{f}, b, \xi$ alone) of (A.2) to be the following:

$$\min_{\eta \in \Delta_{|\mathcal{V}|,1}} \quad \min_{\zeta_v \in \Delta_{|D(v)|,\hat{q}} \forall v \in \mathcal{V}} \quad \max_{\beta_t \in S_m(\mathbf{y}_t, C) \; \forall t \in T} G(\eta, \zeta, \beta) \tag{A.5}$$

where

$$G(\eta, \zeta, \beta) = \sum_{t,i} \beta_{ti} - \frac{1}{2} \sum_{w \in \mathcal{V}} \tau_w(\eta, \zeta) \|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}},$$

$\|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} = \left(\sum_{j=1}^k \left(\beta^\top \mathbf{K}_w^j \beta\right)^{\bar{p}}\right)^{\frac{1}{\bar{p}}}$ and $\bar{p} = \frac{p}{2(p-1)}$
It turns out that at optimality:

$$\Omega_S(\mathbf{f})^2 = \sum_{w \in V} \tau_w(\eta, \zeta) \|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}}, \tag{A.6}$$

which can be realized from the KKT conditions associated with this primal-dual pair.

Note that G is concave in β and convex in η and ζ. Also the feasibility sets involved are convex and compact. Using the Sion-Kakutani minmax theorem (Sion, 1958), we have that the min and max in the above dual can be interchanged (see also Szafranski et al. (2008)) leading to:

$$\max_{\beta \in S(\mathbf{y}, C)} \quad \min_{\eta \in \Delta_{|V|,1}} \quad \min_{\zeta_v \in \Delta_{|D(v)|,\hat{q}} \forall v \in V} G(\eta, \zeta, \beta) \tag{A.7}$$

The duality gap resulting from this interchange will be used (as shown later) for proving Theorem 2. For now, we will concentrate on the second term of equation (A.7), which is:

$$\max_{\eta\in\Delta_{|V|,1}}\max_{\zeta_v\in\Delta_{|D(v)|,\hat{q}}\forall v\in V}\sum_{w\in V}\tau_w(\eta,\zeta)\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}} \tag{A.8}$$

$$\Big|\Big|\Big| \text{ (Lagrange dual wrt. } \eta)$$

$$\max_{\zeta_v\in\Delta_{|D(v)|,\hat{q}}\forall v\in V}\min_{\kappa\in L}\max_{v\in\mathcal{V}}\sum_{w\in D(v)}\frac{\kappa_{vw}^2\zeta_{vw}\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}}{d_v^2} \tag{A.9}$$

$$\Big|\Big|\Big| \text{ (By Sion-Kakutani theorem)}$$

$$\min_{\kappa\in L}\max_{v\in\mathcal{V}}\max_{\zeta_v\in\Delta_{|D(v)|,\hat{q}}}\sum_{w\in D(v)}\frac{\kappa_{vw}^2\zeta_{vw}\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}}{d_v^2} \tag{A.10}$$

$$\Big|\Big|\Big| \text{ (By Lemma (A.1))}$$

$$\min_{\kappa\in L}\max_{v\in\mathcal{V}} d_v^{-2}\left(\sum_{w\in D(v)}\left(\kappa_{vw}^2\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}\right)^{\bar{q}}\right)^{\frac{1}{\bar{q}}} \tag{A.11}$$

where, $L=\left\{\kappa\in\mathbb{R}^{|\mathcal{V}|\times|\mathcal{V}|}\mid \kappa\geq 0,\ \sum_{v\in A(w)}\kappa_{vw}=1,\ \sum_{v\in A(w)^c}\kappa_{vw}=0\ \forall\ w\in\mathcal{V}\right\}$ and $\bar{q}=\frac{q}{2(q-1)}$.

Let us denote the optimal value of (A.11) by $\mathcal{O}$. Consider the following problem which has optimal value $\mathcal{O}^{\bar{q}}$: $\min_{\kappa\in L}\max_{v\in\mathcal{V}} d_v^{-2\bar{q}}\sum_{w\in D(v)}\left(\kappa_{vw}^2\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}\right)^{\bar{q}}$. The Lagrangian is: $\mathcal{L}(\kappa,A,\gamma)=A+\sum_{v\in\mathcal{V}}\gamma_v\left(d_v^{-2\bar{q}}\sum_{w\in D(v)}\left(\kappa_{vw}^2\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}\right)^{\bar{q}}-A\right)$. Minimization of Lagrangian wrt. A leads to the constraint $\gamma\in\Delta_{|\mathcal{V}|,1}$. The minimization problem wrt. κ is given by:

$$\min_{\kappa\in L}\sum_{v\in\mathcal{V}}\sum_{w\in D(v)}\gamma_v\left(d_v^{-2}\kappa_{vw}^2\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}\right)^{\bar{q}}$$

Using the special structure of L, we can re-write this minimization problem as:

$$\sum_{w\in\mathcal{V}}\left(\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}\right)^{\bar{q}}\min_{\kappa_w\in\Delta_{|A(w)|,1}}\sum_{v\in A(w)}\left(\gamma_v d_v^{-2\bar{q}}\right)\kappa_{vw}^{2\bar{q}}$$

Using Holder's inequality[1], we have that:

$$\min_{\kappa_w\in\Delta_{|A(w)|,1}}\sum_{v\in A(w)}\left(\gamma_v d_v^{-2\bar{q}}\right)\kappa_{vw}^{2\bar{q}}=\lambda_w(\gamma)=\left(\sum_{v\in A(w)}d_v^q\gamma_v^{1-q}\right)^{\frac{1}{1-q}}$$

Hence the dual of (A.11) is:

$$\max_{\gamma\in\Delta_{|\mathcal{V}|,1}}\left(\sum_{w\in\mathcal{V}}\lambda_w(\gamma)\left(\|\beta^\top\mathbf{K}_w\beta\|_{\bar{p}}\right)^{\bar{q}}\right)^{\frac{1}{\bar{q}}}=\max_{\gamma\in\Delta_{|\mathcal{V}|,1}}\left(\sum_{w\in\mathcal{V}}\lambda_w(\gamma)\left(\sum_j\left(\beta^\top\mathbf{K}_w^j\beta\right)^{\bar{p}}\right)^{\frac{\bar{q}}{\bar{p}}}\right)^{\frac{1}{\bar{q}}} \tag{A.12}$$

Using the above result in (A.5) and again interchanging the min-max leads to the result in the final specialized partial dual given in theorem 1[2].

This completes the derivation of the partial dual given in theorem 1.

[1]Take vectors $\mathbf{u}_1$ and $\mathbf{u}_2$ as those with entries $d_v^{-1}\gamma_v^{\frac{1}{2\bar{q}}}\kappa_{vw}$ and $d_v\gamma_v^{-\frac{1}{2\bar{q}}}$ for all $v\in A(w)$ respectively. The result follows from the Holder's inequality: $\mathbf{u}_1^\top\mathbf{u}_2\leq\|\mathbf{u}_1\|_{2\bar{q}}\|\mathbf{u}_2\|_q$

[2]From the proof it is clear that if β is obtained by solving the dual (4), then it also is an optimal β in the context of (A.5) and hence can be directly employed for verifying the sufficiency condition

B Derivation of sufficiency condition in Theorem 2

Looking back at the min-max exchange employed to get (A.7) from (A.5), the duality gap with a given triplet (η, ζ, β) is given by (here, $(\mathbf{f}, b, \xi)$ represents the associated primal solution):

$$
\begin{aligned}
= \quad & \max_{\hat{\beta} \in S(\mathbf{y}, C)} G(\eta, \zeta, \hat{\beta}) - \min_{\hat{\eta} \in \Delta_{|V|,1}} \quad \min_{\hat{\zeta}_v \in \Delta_{|D(v)|,\bar{q}} \forall v \in V} G(\hat{\eta}, \hat{\zeta}, \beta) \\
\leq \quad & \tfrac{1}{2}\Omega_S(\mathbf{f})^2 + C \textstyle\sum_{t,i} \xi_{ti} - \min_{\hat{\eta} \in \Delta_{|V|,1}} \quad \min_{\hat{\zeta}_v \in \Delta_{|D(v)|,\bar{q}} \forall v \in V} G(\hat{\eta}, \hat{\zeta}, \beta) \\
= \quad & \underbrace{\Omega_S(\mathbf{f})^2 + C \textstyle\sum_{t,i} \xi_{ti} - \sum_{t,i} \beta_{ti}}_{\text{Gap in solving the problem with fixed } (\eta,\zeta)} + \tfrac{1}{2} \underbrace{\left(\max_{\hat{\eta} \in \Delta_{|V|,1}} \max_{\hat{\zeta}_v \in \Delta_{|D(v)|,\bar{q}} \forall v \in V} \textstyle\sum_{w \in V} \tau_w(\hat{\eta}, \hat{\zeta}) \|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} - \Omega_S(\mathbf{f})^2 \right)}_{\text{Gap in solving the problem with fixed } \beta}
\end{aligned}
$$

Let the optimal solution of the dual formulation restricted to active set $\mathcal{W}$ be $(\eta_\mathcal{W}, \zeta_\mathcal{W}, \beta_\mathcal{W})$. Using the above upper bound and (A.6), it is easy to see that the following condition is sufficient for $(\eta_\mathcal{W}, \zeta_\mathcal{W}, \beta_\mathcal{W})$ having a duality gap less than ϵ w.r.t the original primal formulation (A.2):

$$
\max_{\eta \in \Delta_{|V|,1}} \max_{\zeta_v \in \Delta_{|D(v)|,\bar{q}} \forall v \in V} \sum_{w \in V} \tau_w(\eta, \zeta) \|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}} \leq \sum_{w \in V} \tau_w(\eta_\mathcal{W}, \zeta_\mathcal{W}) \|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}} + 2\epsilon \tag{B.1}
$$

In the following we obtain an upper bound on the L.H.S. term of (B.1) which lead to the sufficiency condition (6) given in Theorem 2. We first note that the L.H.S. term of (B.1) is same as (A.8) and hence is equivalent to (A.11). Since (A.11) is a minimization over κ, any feasible κ for (A.11) will provide us an upper bound on the L.H.S. term. More specifically: for all $w \in \mathcal{W}$ we simply take it to be the optimal κ obtained by solving (A.2) restricted to the active set $\mathcal{W}$. This is fine because $\mathcal{W} = hull(\mathcal{W})$. For all $w \in \mathcal{W}^c$, motivated by the choice in case of $q = 2$ by Bach (2009) (section A.5), we choose it to be: $\kappa_{vw} = d_v \left(\sum_{u \in A(v) \cap \mathcal{W}^c} d_u \right)^{-1}$. With this choice we have that:

$$
\begin{aligned}
& \max_{\eta \in \Delta_{|V|,1}} \max_{\zeta_v \in \Delta_{|D(v)|,\bar{q}} \forall v \in V} \sum_{w \in V} \tau_w(\eta, \zeta) \|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} \\
\leq \quad & (\because \text{Specific choice of } \kappa) \\
& \max \left\{ \Omega_S(\mathbf{f}_\mathcal{W})^2, \max_{t \in \mathcal{W}^c} \left(\sum_{w \in D(t)} \left(\frac{\|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}}}{\left(\sum_{v \in A(w) \cap \mathcal{W}^c} d_v \right)^2} \right)^{\bar{q}} \right)^{\frac{1}{\bar{q}}} \right\} \\
= \quad & (\because \mathcal{W} = hull(\mathcal{W})) \\
& \max \left\{ \Omega_S(\mathbf{f}_\mathcal{W})^2, \max_{t \in sources(\mathcal{W}^c)} \left(\sum_{w \in D(t)} \left(\frac{\|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}}}{\left(\sum_{v \in A(w) \cap \mathcal{W}^c} d_v \right)^2} \right)^{\bar{q}} \right)^{\frac{1}{\bar{q}}} \right\} \\
\leq \quad & (\because \textstyle\sum_{v \in A(w) \cap \mathcal{W}^c} d_v \geq \sum_{v \in A(w) \cap D(t)} d_v) \\
& \max \left\{ \Omega_S(\mathbf{f}_\mathcal{W})^2, \max_{t \in sources(\mathcal{W}^c)} \left(\sum_{w \in D(t)} \left(\frac{\|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}}}{\left(\sum_{v \in A(w) \cap D(t)} d_v \right)^2} \right)^{\bar{q}} \right)^{\frac{1}{\bar{q}}} \right\}
\end{aligned}
$$

Employing this upper bound in (B.1), we get:

$$
\max_{t \in sources(\mathcal{W}^c)} \left(\sum_{w \in D(t)} \left(\frac{\|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}}}{\left(\sum_{v \in A(w) \cap D(t)} d_v \right)^2} \right)^{\bar{q}} \right)^{\frac{1}{\bar{q}}} \leq \sum_{w \in V} \tau_w(\eta_\mathcal{W}, \zeta_\mathcal{W}) \|\beta_\mathcal{W}^\top \mathbf{K}_w \beta_\mathcal{W}\|_{\bar{p}} + 2\epsilon \tag{B.2}
$$

Now we know that for any vector $\mathbf{z}$, $\|\mathbf{z}\|_\rho \leq \|\mathbf{z}\|_1 \ \forall \rho \geq 1$. This implies $\|\beta_{\mathcal{W}}^\top \mathbf{K}_w \beta_{\mathcal{W}}\|_{\bar{p}} \leq \sum_j \beta_{\mathcal{W}}^\top \mathbf{K}_w^j \beta_{\mathcal{W}}$.[3] Using the same inequality again for the $\bar{q}$-norm present in (B.2), we get the following sufficiency condition:

$$\max_{t \in sources(\mathcal{W}^c)} \left(\sum_{w \in D(t)} \frac{\sum_j \beta_{\mathcal{W}}^\top \mathbf{K}_w \beta_{\mathcal{W}}}{\left(\sum_{v \in A(w) \cap D(t)} d_v\right)^2} \right) \leq \sum_{w \in V} \tau_w(\eta_{\mathcal{W}}, \zeta_{\mathcal{W}}) \|\beta_{\mathcal{W}}^\top \mathbf{K}_w \beta_{\mathcal{W}}\|_{\bar{p}} + 2\epsilon \tag{B.3}$$

From the proof of Theorem 1, we had obtained the following equality:

$$\max_{\eta \in \Delta_{|V|,1}} \max_{\zeta_v \in \Delta_{|D(v)|,\bar{q}} \forall v \in V} \sum_{w \in V} \tau_w(\eta, \zeta) \|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} = \max_{\gamma \in \Delta_{|\mathcal{V}|,1}} \left(\sum_{w \in \mathcal{V}} \lambda_w(\gamma) \left(\sum_j \left(\beta^\top \mathbf{K}_w^j \beta\right)^{\bar{p}} \right)^{\frac{\bar{q}}{\bar{p}}} \right)^{\frac{1}{\bar{q}}} \tag{B.4}$$

It can easily be seen that at optimality, the equality also holds when the formulation is restricted to the active set $\mathcal{W}$ (since terms corresponding to nodes outside the active set will be zero in both L.H.S and R.H.S.). Let $(\hat{\gamma}, \hat{\beta})$ be the optimal solution of (4) restricted to $\mathcal{W}$. It is clear from the above derivations that $\beta_{\mathcal{W}} = \hat{\beta}$ and hence the following equality holds:

$$\sum_{w \in V} \tau_w(\eta_{\mathcal{W}}, \zeta_{\mathcal{W}}) \|\beta_{\mathcal{W}}^\top \mathbf{K}_w \beta_{\mathcal{W}}\|_{\bar{p}} = \left(\sum_{w \in \mathcal{V}} \lambda_w(\hat{\gamma}) \left(\sum_j \left(\hat{\beta}^\top \mathbf{K}_w^j \hat{\beta}\right)^{\bar{p}} \right)^{\frac{\bar{q}}{\bar{p}}} \right)^{\frac{1}{\bar{q}}}$$

By substituting the above result in (B.3), we arrive at the final sufficiency condition in Theorem 2.
This completes the proof of Theorem 2.

C Applicability of Mirror-Descent Algorithm

For the applicability of mirror descent algorithm in our case, we need to prove that H is Lipschitz continuous. We prove the following statement: if all the eigen-values of the gram-matrices $\mathbf{K}_w^j$ are finite and non-zero, then H is Lipschitz continuous.
We begin by recalling the expression for the i^{th} entry in the sub-gradient

$$(\nabla H(\gamma))_i = \underbrace{-\frac{1}{2\bar{q}}}_{T_1} \underbrace{\left(\sum_{w \in \mathcal{V}} \lambda_w(\gamma) \left(\|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} \right)^{\bar{q}} \right)^{\frac{1}{\bar{q}} - 1}}_{T_2} \underbrace{\left(\sum_{w \in D(i)} d_i^q \gamma_i^{-q} \lambda_w(\gamma)^q \left(\|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} \right)^{\bar{q}} \right)}_{T_3}$$

We show that g is Lipschitz continuous by showing that the sub-gradient is bounded. Since $\bar{q} \in [1, \infty)$, we have that $|T_1| \leq 1/2$. Let the maximum and minimum eigenvalues over all $\mathbf{K}_w^j$ be α, π respectively. Then, we have that $n\alpha \|\bar{\beta}\|^2 \leq \|\beta^\top \mathbf{K}_w \beta\|_{\bar{p}} \leq n\pi \|\bar{\beta}\|^2$, where n is the number of given base kernels. Using this, we obtain: $\sum_{w \in \mathcal{V}} \lambda_w(\gamma) \left(\|\bar{\beta}^\top \mathbf{K}_w \bar{\beta}\|_{\bar{p}} \right)^{\bar{q}} \geq (n\alpha)^{\bar{q}} \|\bar{\beta}\|^{2\bar{q}} \sum_{w \in \mathcal{V}} \lambda_w(\gamma)$. Note that λ_w is zero whenever any $\gamma_v = 0, v \in A(w)$. Hence, by (A.12) it is clear that the set of non-zero γ is such that it is equal to its hull. Also, since $\gamma \in \Delta_{|V|,1}$, we must have that atleast one $\gamma_u \geq \frac{1}{|V|}$. This gives us that $\sum_{w \in \mathcal{V}} \lambda_w(\gamma) \geq d_{max}^{q/(1-q)}/|V|$ where d_{max} is the maximum of $d_v, v \in \mathcal{V}$. Note that $\frac{1}{\bar{q}} - 1 \leq 0 \ \forall q \in (1, 2]$. Thus we obtain: $T_2 \leq \left(n\alpha \|\bar{\beta}\|^2 / |V|^{1/\bar{q}}\right)^{1-\bar{q}} d_{max}^{\frac{q-2}{1-q}}$. Now, it is easy to see that: $\gamma_i^{-q} \lambda_w(\gamma)^q \leq d_{min}^{q^2/(1-q)}$. Hence $T_3 \leq |V| (n\pi)^{\bar{q}} \|\bar{\beta}\|^{2\bar{q}} d_{max}^q d_{min}^{q^2/(1-q)}$. Also, because $0 \leq \bar{\beta} \leq C$, we have that $\|\bar{\beta}\| \leq \sqrt{m}C$. Summarizing these finding we obtain the following bound on the sub-gradient:

$$\|\nabla H(\gamma)\|_1 \leq \frac{mC^2}{2} n \alpha^{1-\bar{q}} \pi^{\bar{q}} |V|^{\frac{2}{q}} d_{max}^{\frac{q-2}{1-q}+q} d_{min}^{\frac{q^2}{1-q}}$$

[3] This is because $\mathbf{K}_w^j$ are positive semi-definite kernel matrices and hence $\Rightarrow \mathbf{z}^\top \mathbf{K}_w^j \mathbf{z} >= 0 \ \forall \mathbf{z} \in \mathbb{R}^{mT}$. Therefore, $\|\beta_{\mathcal{W}}^\top \mathbf{K}_w \beta_{\mathcal{W}}\|_1 = \sum_j \beta_{\mathcal{W}}^\top \mathbf{K}_w^j \beta_{\mathcal{W}}$.

References

Bach, F. High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning. Technical report, INRIA, France, 2009.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, 2004.

Evgeniou, T. and Pontil, M. Regularized multi–task learning. In *ACM SIGKDD*, 2004.

Micchelli, Charles A. and Pontil, Massimiliano. Learning the Kernel Function via Regularization. *JMLR*, 6:1099–1125, 2005.

Rakotomamonjy, Alain, Bach, Francis, Canu, Stéphane, and Grandvalet, Yves. Simplemkl. *Journal of Machine Learning Research*, 9:2491–2521, 11 2008.

Schölkopf, Bernhard and Smola, Alex. *Learning with Kernels*. MIT press, Cambridge, 2002.

Sion, M. On General Minimax Theorem. *Pacific Journal of Mathematics*, 1958.

Szafranski, M., Grandvalet, Y., and Rakotomamonjy, A. Composite Kernel Learning. In *ICML*, 2008.