Supplemental Materials for "Learning Kernels with Radiuses of Minimum Enclosing Balls"

Abstract

In this document, we first give an example where the radius of the uniform combination of basis kernels is far smaller than the radius of each basis kernel. Then, we present proofs for Proposition 2, Theorem 1 and Proposition 4. Finally, we illustrate the convergence speed of our algorithm by experiments.

1 An example

In the linear combination case $k^{(\theta)} = \sum_j \theta_j k_j$ ($\theta_j > 0$), consider the squared radius of MEB in the feature space endowed with $k^{(\theta)}$, which is denoted by $r^2(\theta) \doteq R^2(k^{(\theta)})$. The convexity of $r^2(\theta)$ is stated in the following.

Lemma 1. *In the linear combination case $k^{(\theta)} = \sum_j \theta_j k_j$ ($\theta_j > 0$), the squared radius $r^2(\theta)$ of MEB in the feature space endowed with $k^{(\theta)}$ is a convex function w.r.t. θ.*

Proof. In the linear combination case, the squared radius is equal to

$$r^2(\theta) = \min_{y,c_j} \quad y, \quad \text{s.t.} \quad \sum_j \|\sqrt{\theta_j}\phi(x_i; k_j) - c_j\|^2 - y \leq 0. \tag{1}$$

The distance term in the constraint is equal to: $\sum_j \|\sqrt{\theta_j}\phi(x_i; k_j) - c_j\|^2 = \sum_j \big(\theta_j k_j(x_i, x_i) + \|c_j\|^2 - 2\langle\sqrt{\theta_j}\phi(x_i; k_j), c_j\rangle\big)$. Let $\tilde{c_j} = \sqrt{\theta_j}c_j$, then we have

$$r^2(\theta) = \min_{y,\tilde{c_j}} \; y, \quad \text{s.t.} \; \sum_j \big(\theta_j k_j(x_i, x_i) + \|\tilde{c_j}\|^2/\theta_j - 2\langle\phi(x_i; k_j), \tilde{c_j}\rangle\big) - y \leq 0. \tag{2}$$

Note the objective function is convex w.r.t. $\{\theta, \tilde{c_j}, y\}$, and the constraint is also convex w.r.t. $\{\theta, \tilde{c_j}, y\}$ due to the convexity of $\|\tilde{c_j}\|^2/\theta_j$ (which is proven in [1]). Thus, $r^2(\theta)$, which is the partial minimal value of the convex problem (2) w.r.t. only $\{\tilde{c_j}, y\}$, is convex w.r.t. θ. □

Now we further suppose $R(k_j)$, which are the radiuses of basis kernels k_j, are all equal to B, and also impose an L_1 norm constraint $\sum_j \theta_j = 1$. Due to the convexity, for any θ, we have $r^2(\theta) \leq B^2$. This shows how the L_1 norm constraint guarantees an upper bound on $r^2(\theta)$. However, this bound may be very loose. For example, let p kernel matrices $K^j \doteq [k_j(x_a, x_b)]_{ab} = {v_j}^\top v_j$, where v^j is a vector of length $2p$, the $(2j-1)$-th and the $(2j)$-th elements of v_j is B and $-B$, respectively, and other elements are zero. Then the squared radius of each K_j is equal to B^2, whereas the squared radius of their uniform combination $K_{\text{unif}} = \sum_{j=1}^p \frac{1}{p} K^j$ is no larger than the maximal diagonal element of K_{unif}, which is equal to $\frac{B^2}{p}$ (the maximal diagonal element corresponds the squared radius of the enclosing ball centered at the origin, which is no smaller than MEB). This example shows that for large p, the radius of the learned kernel under the L_1 norm constraint may be far smaller than the upper bound B. With such a loose bound on $r^2(\theta)$, minimizing $\|w\|^2$ itself gives no guarantee for obtaining the kernel with the smallest (or approximately smallest) $r^2(\theta)\|w\|^2$. Instead of using a loose upper bound, our approach exactly handle the radius of MEB of the learned kernel.

2 Proof of Proposition 2

Proposition 2. *Given any norm definition $\mathcal{N}(\cdot)$ and any set $\mathcal{S} \subseteq \mathbb{R}$, suppose there exists $c > 0$ that satisfies $c \in \mathcal{S}$. Let (a) denote the problem of minimizing $g_{\text{linear}}(\theta)$ s.t. $\theta_i \geq 0$, and (b) denote the problem of minimizing $g_{\text{linear}}(\theta)$ s.t. $\theta_i \geq 0$ and $\mathcal{N}(\theta) \in \mathcal{S}$. Then we have: (1) For any local (global) optimal solution of (a), denoted by θ^a, $\frac{c}{\mathcal{N}(\theta^a)}\theta^a$ is also the local (global) optimal solution of (b). (2) For any local (global) optimal solution of (b), denoted by θ^b, θ^b is also the local (global) optimal solution of (a).*

Proof. For conclusion (1), as θ^a is the local optimal solution of (a), there exists $\delta > 0$ that for any θ^1 ($\theta^1_i \geq 0$) that satisfies $\|\theta^1 - \theta^a\| \leq \frac{\mathcal{N}(\theta^a)}{c}\delta$, we have $g(\theta^1) \geq g_{\text{linear}}(\theta^a) = g_{\text{linear}}(\frac{c}{\mathcal{N}(\theta^a)}\theta^a)$. Then, for any θ ($\theta_i \geq 0$) that satisfies $\|\theta - \frac{c}{\mathcal{N}(\theta^a)}\theta^a\| \leq \delta$, we have $\|\frac{\mathcal{N}(\theta^a)}{c}\theta - \theta^a\| = \frac{\mathcal{N}(\theta^a)}{c}\|\theta - \frac{c}{\mathcal{N}(\theta^a)}\theta^a\| \leq \frac{\mathcal{N}(\theta^a)}{c}\delta$, and thus $g_{\text{linear}}(\frac{\mathcal{N}(\theta^a)}{c}\theta) \geq g_{\text{linear}}(\frac{c}{\mathcal{N}(\theta^a)}\theta^a)$, and then $g_{\text{linear}}(\theta) = g_{\text{linear}}(\frac{\mathcal{N}(\theta^a)}{c}\theta) \geq g_{\text{linear}}(\frac{c}{\mathcal{N}(\theta^a)}\theta^a)$. Due to $\mathcal{N}(\frac{c}{\mathcal{N}(\theta^a)}\theta^a) = c \in \mathcal{S}$, $\frac{c}{\mathcal{N}(\theta^a)}\theta^a$ also satisfies the constraint of (b), and thus $\frac{c}{\mathcal{N}(\theta^a)}\theta^a$ is the local optimal solution of (b). If θ^a is the global optimal solution of (a), we can set δ to be ∞, and then $\frac{c}{\mathcal{N}(\theta^a)}\theta^a$ is also the global solution of (b).

For proof of conclusion (2), we use two types of distances: $d_1(x, y) = \mathcal{N}(x - y)$ and $d_2(x, y) = |\mathcal{N}(x) - \mathcal{N}(y)| + \mathcal{N}(x - y)$ (It is easy to verify that d_1 and d_2 satisfy the metric conditions). On one hand, since θ^b is the local optimal solution of (b), there exists $\delta > 0$ that for any θ^2 ($\theta^2_i \geq 0$) that satisfies $d_1(\theta^2, \theta^b) \leq \delta$ and $\mathcal{N}(\theta^2) \in \mathcal{S}$, we have $g_{\text{linear}}(\theta^b) \leq g_{\text{linear}}(\theta^2)$. On the other hand, for any θ ($\theta_i \geq 0$) that satisfies $d_2(\theta, \theta^b) \leq \delta$, we have $\mathcal{N}(\frac{\mathcal{N}(\theta^b)}{\mathcal{N}(\theta)}\theta) = \mathcal{N}(\theta^b) \in \mathcal{S}$ and $d_1(\frac{\mathcal{N}(\theta^b)}{\mathcal{N}(\theta)}\theta, \theta^b) = \mathcal{N}(\frac{\mathcal{N}(\theta^b)}{\mathcal{N}(\theta)}\theta - \theta + \theta - \theta^b) \leq \mathcal{N}(\frac{\mathcal{N}(\theta^b)}{\mathcal{N}(\theta)}\theta - \theta) + \mathcal{N}(\theta - \theta^b) = d_2(\theta, \theta^b) \leq \delta$. Note $\frac{\mathcal{N}(\theta^b)}{\mathcal{N}(\theta)}\theta$ satisfies the conditions of θ^2, and thus we have $g(\frac{\mathcal{N}(\theta^b)}{\mathcal{N}(\theta)}\theta) \geq g_{\text{linear}}(\theta^b)$. Due to the scaling invariance, we have $g(\theta) \geq g_{\text{linear}}(\theta^b)$ (for any θ that meets $d_2(\theta, \theta^b) \leq \delta$). Therefore, θ^b is also the local optimal solution of (b). If θ^b is the global optimal solution of (b), let δ be ∞, and then θ^b is also the global optimal solution of (a). □

3 Proof of Theorem 1

First we generalize Danskin's theorem [2], as the following.

Lemma 2. *Let X be a metric space and U be a normed space. Suppose that for all $x \in X$ the function $f(x, \cdot)$ is differentiable, that $f(x, u)$ and $\frac{\partial f(x,u)}{\partial u}$ are continuous on $X \times U$ and let Φ be a compact subset of X. Let define the optimal value function as $v(u) = \inf_{x \in \Phi} f(x, u)$. (a) The optimal value function is directionally differentiable. Furthermore, if for any $u \in U$, $f(\cdot, u)$ has a unique minimizer x^* over Φ then (b1) $v(u)$ is differentiable, (b2) $\frac{dv(u)}{du} = \frac{\partial f(x^*,u)}{\partial u}$, (b3) the minimizer $x^*(u)$ is continuous w.r.t. u, and (b4) $\frac{dv(u)}{du}$ is also continuous.*

Proof. Statements (a) (b1) and (b2) have been proven by Danskin [2]. Below we give the proofs of conclusions (b3) and (b4).

First we prove that the minimizer $x^*(u)$ is continuous w.r.t. u by contradiction. Given u^0 and the corresponding minimizer $x^0 = x^*(u^0)$, let us assume there exists a sequence $u^m \to u^0$, and the the corresponding minimizers $x^m = x^*(u^m)$ does not converge to x^0. Then there must exist a positive value $\epsilon > 0$ so that for any large number N, there exist $i > N$ and $\|x^i - x^0\| > \epsilon$. Thus we can pick up an infinite subsequence $x^{m'}$, which satisfies $\|x^{m'} - x^0\| > \epsilon$. Because Φ is compact, there must exist a converging sub-subsequence $x^{m''}$ in $x^{m'}$ and a limit s so that $x^{m''} \to s$. From $\|x^{m''} - x^0\| > \epsilon$ we get conclusion (A): $\|s - x^0\| \geq \epsilon$.

On the other hand, consider the corresponding sequence $u^{m''}$. Because the $x^{m''}$ is the minimizer of $f(\cdot, u^{m''})$, we have $f(x^{m''}, u^{m''}) \leq f(x^0, u^{m''})$. Because f is continuous, $x^{m''} \to s$ and $u^{m''} \to u^0$, we get $f(s, u^0) \leq f(x^0, u^0)$. As x^0 is the unique minimizer, we get conclusion (B): $s = x^0$.

Note that (A) and (B) are contradict, and thus we finish the proof that $x^*(u)$ is continuous (Conclusion (b3)).

Conclusion (b2) states the derivative $\frac{dv(u)}{du} = \frac{\partial f(x^*,u)}{\partial u}$. Due to the continuity of $x^*(\cdot)$ and $\frac{\partial f(\cdot,\cdot)}{\partial u}$, we get: $\frac{dv(u)}{du}$ is continuous (Conclusion (b4)). □

Then, by use of the above lemma, we get the following theorem.

Theorem 1. *Let Y be a metric space, X, U and Z be normed spaces. Suppose that: (1) for all $x \in X$ the function $g_1(x,\cdot,\cdot)$ is differentiable, (2) the function $g_1(x,u,z)$, $\frac{\partial g_1(x,u,z)}{\partial u}$ and $\frac{\partial g_1(x,u,z)}{\partial z}$ are continuous on $X \times U \times Z$, (3) for all $y \in Y$ the function $g_2(y,\cdot,\cdot)$ ($g_2 : Y \times X \times U \to Z$) is differentiable, (4) the function $g_2(y,x,u)$, $\frac{\partial g_2(y,x,u)}{\partial x}$ and $\frac{\partial g_2(y,x,u)}{\partial u}$ are continuous on $Y \times X \times U$, (5) sets Φ_X and Φ_Y are compact subsets of X and Y, respectively.*

Let us define a bi-level optimal value function as

$$v_1(u) = \inf_{x\in\Phi_X} g_1(x,u,v_2(x,u)), \tag{3}$$

where $v_2(x,u)$ is another optimal value function as

$$v_2(x,u) = \inf_{y\in\Phi_Y} g_2(y,x,u). \tag{4}$$

If for any x and u, $g_2(\cdot,x,u)$ has a unique minimizer $y^*(x,u)$ over Φ_Y, then $y^*(x,u)$ are continuous on $X \times U$, and $v_1(u)$ is directionally differentiable. Furthermore, if for any u, the $g_1(\cdot,u,v_2(\cdot,u))$ has also a unique minimizer $x^*(u)$ over Φ_X, then

1. *the minimizer $x^*(u)$ are continuous on U,*
2. *$v_1(u)$ is continuously differentiable, and its derivative is equal to*

$$\frac{dv_1(u)}{du} = \left(\frac{\partial g_1(x^*,u,v_2)}{\partial u} + \frac{\partial v_2(x^*,u)}{\partial u}\frac{\partial g_1(x^*,u,v_2)}{\partial v_2}\right)\Big|_{v_2=v_2(x^*,u)}, \text{ where } \frac{\partial v_2(x^*,u)}{\partial u} = \frac{\partial g_2(y^*,x^*,u)}{\partial u}. \tag{5}$$

Proof. Because $g_2(y,x,u)$ is continuously differentiable w.r.t. x and u, using conditions (3)-(5) and Lemma 2 we get: (I) $y^*(x,u)$ are continuous on $X \times U$, and (II) $v_2(x,u)$ is continuously differentiable w.r.t. x and u, and

$$\frac{\partial v_2(x,u)}{\partial u} = \frac{\partial g_2(y^*,x,u)}{\partial u}. \tag{6}$$

Since $g_1(x,u,v_2)$ is continuous w.r.t. x, u and v_2, and $v_2(x,u)$ is continuous w.r.t. x and u, thus $g_1(x,u,v_2(x,u))$ is continuous w.r.t. x and u. Then, because $g_1(x,u,v_2)$ is continuously differentiable w.r.t. u and v_2 and $v_2(x,u)$ is continuously differentiable w.r.t. u, we get that $g_2(x,u,v_2(x,u))$ is continuously differentiable w.r.t. u. Using Lemma 2, we get: $v_1(u)$ is directionally differentiable.

If for any u, the $g_1(\cdot,u,v_2(\cdot,u))$ has also a unique minimizer $x^*(u)$ over Φ_X, using Lemma 2 we have:

1. $v_1(u)$ is differentiable,
2. the derivative is equal to

$$\frac{dv_1(u)}{du} = \left(\frac{\partial g_1(x^*,u,v_2)}{\partial u} + \frac{\partial v_2(x^*,u)}{\partial u}\frac{\partial g_1(x^*,u,v_2)}{\partial v_2}\right)\Big|_{v_2=v_2(x^*,u)}, \text{ where } \frac{\partial v_2(x^*,u)}{\partial u} = \frac{\partial g_2(y^*,x^*,u)}{\partial u}. \tag{7}$$

3. the minimizer $x^*(u)$ and the derivative $\frac{dv_1(u)}{du}$ are continuous w.r.t. u.

□

Remarkably, the bi-level optimization problem defined in (3) is a more general form than the min-max problem. Give any min-max problem:

$$o(u) = \min_x \max_y g_3(x,y,u). \tag{8}$$

Set $g_1(x,u,v_2) = -v_2$ and $g_2(y,x,u) = -g_3(x,y,u)$. Then the bi-level optimization problem (3) is equivalent to the min-max problem: $v_1(u) = o(u)$. On the inverse direction, when $g_1(x,u,v_2)$ is not monotone to v_2, the bi-level optimization problem can not be transformed to be a min-max problem.

4 Proof of Proposition 4

Proposition 4. *In linear combination cases, for any local optimal solution of the RKL problem, denoted by θ^*, there exist $C_1 > 0$ and $C_2 > 0$ that θ^* is the global optimal solution of the following convex problem:*

$$\min_{\theta,w_j,b,\xi_i} \tfrac{1}{2}\sum_j \|w_j\|^2 + C_1 r^2(\theta) + C_2 \sum_i \xi_i^2,\ \text{s.t. } y_i(\textstyle\sum_j \langle w_j, \phi(x_i;\theta_j k_j)\rangle + b) + \xi_i \geq 1,\ \xi_i \geq 0. \quad (9)$$

Proof. On one hand, in the linear combination cases the RKL problem is as

$$\min_{\theta,w_j,b,\xi_i} \tfrac{1}{2}\sum_j \|w_j\|^2 r^2(\theta) + C \sum_i \xi_i^2,\ \text{s.t. } y_i(\textstyle\sum_j \langle w_j, \phi(x_i;\theta_j k_j)\rangle + b) + \xi_i \geq 1,\ \xi_i \geq 0. \quad (10)$$

Let $\frac{\tilde{w}_j}{\sqrt{\theta_j}} = w_j$ to substitute w_j, we get

$$\min_{\theta,w_j,b,\xi_i} \tfrac{1}{2}\sum_j \frac{\tilde{w}_j^2}{\theta_j} r^2(\theta) + C \sum_i \xi_i^2,\ \text{s.t. } y_i(\textstyle\sum_j \langle \tilde{w}_j, \phi(x_i;k_j)\rangle + b) + \xi_i \geq 1,\ \xi_i \geq 0. \quad (11)$$

On the other hand, problem (9) is equivalent to

$$\min_{\theta,w_j,b,\xi_i} \frac{C_3}{2}\sum_j \frac{\tilde{w}_j^2}{\theta_j} + C_1 r^2(\theta) + C_2 \sum_i \xi_i^2,\ \text{s.t. } y_i(\textstyle\sum_j \langle \tilde{w}_j, \phi(x_i;k_j)\rangle + b) + \xi_i \geq 1,\ \xi_i \geq 0, \quad (12)$$

with $C_3 = 1$. Due to the convexity of $r^2(\theta)$ and $\frac{\tilde{w}_j^2}{\theta_j}$, problem (12) has a convex objective function and convex constraints.

Note that the constraints in (11) and (12) are the same, and also the same with the constraints in SVM. It has been well known that for such constraints, the regularity conditions of LICQ (Linear independence constraint qualification) are satisfied if $K(\theta)$ is strictly positive definite (which is also easy to be verified). So the necessary condition of any local optimal solution of (11) is that it satisfies the K.K.T. (Karush-Kuhn-Tucker) conditions. Let $z = \{\theta, \tilde{w}, b\}$, and

$$f_1(z) = \textstyle\sum_j \frac{\tilde{w}_j^2}{\theta_j}, \quad f_2(z) = r^2(\theta), \quad f_3(z) = \sum_i \xi_i^2, \quad (13)$$

$$g_i(z) = 1 - y_i(\textstyle\sum_j \langle \tilde{w}_j, \phi(x_i;k_j)\rangle + b) - \xi_i, \quad g_{n+i} = -\xi_i. \quad (14)$$

For any local optimal solution of (11), denoted by z^*, it must satisfies the K.K.T. conditions, as:

$$\tfrac{1}{2} f_2(z^*)\tfrac{df_1(z^*)}{dz^*} + \tfrac{1}{2} f_1(z^*)\tfrac{df_2(z^*)}{dz^*} + C\tfrac{df_3(z^*)}{dz^*} + \textstyle\sum_{i=1}^{2n} \mu_i \frac{dg_i(z^*)}{dz^*} = 0, \quad (15)$$

$$g_i(z^*) \leq 0, \quad \mu_i \geq 0, \quad \mu_i g_i(z^*) = 0. \quad (16)$$

Note this is also the K.K.T. conditions of problem (12) when $C_1 = \frac{1}{2} f_1(z^*)$, $C_2 = C$ and $C_3 = f_2(z^*)$. Since problem (12) is a (continuously differentiable) convex minimization problem, z^* is the global optimal solutions of (12) with above settings of C_i. Problem (12) with these settings of C_i is also equivalent to problem (12) with new settings $C_1 = \frac{f_1(z^*)}{2f_2(z^*)}$, $C_2 = \frac{C}{f_2(z^*)}$ and $C_3 = 1$, which is again equivalent to problem (9). So, for any local solution θ^* of the RKL problem, there exist $C_1 > 0$ and $C_2 > 0$ that θ^* is the global optimal solution of problem (9). □

5 Convergence speed

Here we illustrate the computational efficiency of our presented algorithm. Figure 1 shows the objective functions along with invocation numbers of SVM and MEB solvers on the Ionosphere set and the Splice set. Different types of norm constraints are used for comparison: L_1, L_2 and no norm constraint. As stated, the RKL formulations with different types of norm constraints are equivalent to each other, and thereby they result in the same optimal values. All three lines in the left figure on the Ionosphere set show rapid convergence speeds: the objective function decreases very rapidly, and get an approximate convergence within 10 invocations of SVM and MEB solvers. The right figure on the Splice set gives similar results, where the objective values also approximately converge within 10 invocations.

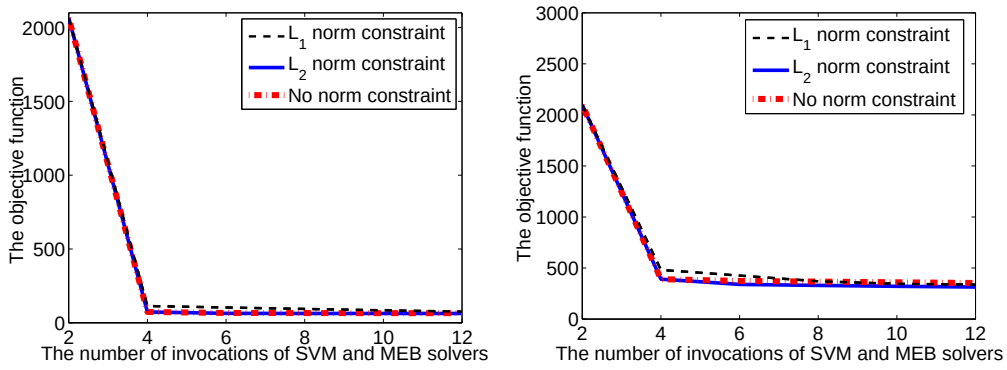

Figure 1: Convergence speeds of RKL with $C = 100$. Left: on the Ionosphere set. Right: on the Splice set.

References

[1] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

[2] J.M. Danskin. The theory of max-min, with applications. *SIAM Journal on Applied Mathematics*, pages 641–664, 1966.