A Sequential Minimal Optimization for the Principle of Relevant Information

Luis G. Sánchez Giraldo
Computational Neuro-Engineering Laboratory
Department of Electrical and Computer Engineering
University of Florida
Gainesville, FL 32611
sanchez@cnel.ufl.edu

Abstract

In this report, we present a sequential minimal optimization routine for the principle of relevant information. The optimization problem is broken down into smaller subproblems that can be solved sequentially. Conditions under which this approach is feasible are verified and an algorithm is proposed. Although the algorithm has computation complexity can be $\mathcal{O}(n^2)$, the memory complexity is linear in the size of the sample allowing for larger data sets that are common in practical applications.

1 Introduction

Support vector machines and in general kernel methods have become standard in machine learning community. These methods have become widely accepted and have shown good results in several practical applications. However, a great part of their success is due to possibility of implementing routines that provide solutions for moderately large scale problems. Otherwise, these methods would be merely an interesting mental exercise. One of the main technical difficulties when faced with real applications is that the size of the dataset becomes a very important constraint when it comes to selecting a method. The time and memory complexities become relevant for data sets above five thousand instances. In off-line scenarios the most decisive factor is the memory complexity [1] since super-linear growth of storage become intractable for the size of the samples commonly encountered in current and new applications.

In the literature, two main approaches have been taken to make kernel methods applicable in large scale problems. Low rank approximations such as Nyström [12], incomplete Cholesky decomposition [2, 10], subset methods [11], decompose the Gram matrix $\mathbf{K}$ into a product $\mathbf{G}\mathbf{G}^{\mathrm{T}}$ where $\mathbf{G}$ is a $(n \times k)$ matrix with $k \ll n$. This methods scale with computational complexity $\mathcal{O}(nk^2)$ and memory complexity $\mathcal{O}(nk)$. These methods represent significant computational savings when the

[1] under reasonable time complexity (polynomial complexity)

eigenvalues of the Gram matrix decay rapidly; nevertheless, as the dimensionality of the input space increases, we may run into difficulties since the rate of decay in the spectrum may require larger ranks(larger k) to obtain the desired accuracy.

The second approach is related to the exploiting properties of optimization problem such as convexity the type of constraints. This has been the case for algorithms such support vector machines for which, optimization can be broken down into smaller subproblems that can be iteratively solved in order to solve the full problem. Early stages employed chunking [1], which breaks down the problem by discarding vectors with zero Lagrange multipliers from subsets of points, the complexity of the problem is reduced to the number of support vectors. Further analysis on the optimality of solving a sequence of subproblems was provided in [6], from which an algorithm that solves subproblems of fixed size was shown to converge to the global optimum. Keeping the size of the subproblem independent of the number of support vectors is extremely important since even the chunking scheme becomes intractable when the number of support vectors exceed the memory resources. Further developments on the above idea led to the SVM^{light} algorithm described in [3] and the sequential minimal optimization (SMO) proposed in [7]. Althought SVM^{light} has been shown experimentally to be faster, the SMO algorithm is particularly appealing in the sense that there is no need to resort to a quadratic programing routine to solve the subproblems. The solution to the subproblem in SMO can be found analytically since it only involve two variables at the time.

In our work, we develop an optimization scheme similar in spirit with the SMO algorithm to find a solution to an objective function called the principle of relevant information (PRI) [8]. This objective function is motivated by information theory, by considering that the process of finding regularities in the data correspond to a constrained minimization of its entropy. Section 2 introduces information quantities based on Renyi's entropy along with the PRI objective function. Next in Section 3, a constrained optimization problem based on the estimators of the quantities involved in the PRI objective function is derived. We prove that solving the first order KKT conditions is necessary and sufficient for optimality. In Section 4, the conditions for which the decomposition of the problem into subproblems guarantee convergence, are verified. Following, a detailed presentation of the derivation of a sequential minimal optimization for the PRI is provided along with some considerations for implementation. Results in Section 6 focus on observing the behavior of the algorithm in terms of computation time on large samples and different regimes of operation. Finally, some conclusions and a motivation for future work are discussed in Section 7.

2 Elements from Renyi's Entropy and and Relevant Information

Renyi's α-order entropy entropy is a natural extension of the widely known Shannon's entropy [9]. In the continuous case for a random variable X with probability density function (PDF) $f(x)$ and support $\mathcal{X}$, the α-entropy $H_\alpha(X)$ is defined as:

$$H_\alpha(f) = \frac{1}{1-\alpha} \log \int_{\mathcal{X}} f^\alpha(x) \mathrm{d}x. \tag{1}$$

As $\alpha \to 1$ we approximate to Shannon's entropy. Extensions for relative entropy also exist, a modified version of Renyi's definition of α-relative entropy between random variables with PDFs f and g is given in [4],

$$D_\alpha(f\|g) = \log \frac{\left(\int g^{\alpha-1} f\right)^{\frac{1}{1-\alpha}} \left(\int g^\alpha\right)^{\frac{1}{\alpha}}}{\left(\int f^\alpha\right)^{\frac{1}{\alpha(1-\alpha)}}}. \tag{2}$$

Similarly, Shannon's relative entropy (Kullback-Leibler divergence) is the limit for $\alpha \to 1$. An important component in the relative entropy is the cross-entropy term $H_\alpha(f; g)$ that quantifies the information gain from observing g with respect to the "true" density f. It turns out that for the case of $\alpha = 2$, the above quantities can be expressed, under some restrictions, as functions of inner products between PDFs. In particular, the 2-order entropy of f and cross-entropy between f and g, can be respectively expressed as,

$$\begin{aligned} H_2(f) &= -\log \int_{\mathcal{X}} f^2(x)\mathrm{d}x; \\ H_2(f; g) &= -\log \int_{\mathcal{X}} f(x)g(x)\mathrm{d}x, \end{aligned} \tag{3}$$

the associated relative entropy of order 2 is called the Cauchy-Schwarz divergence and is defined as follows:

$$D_{\text{CS}}(f\|g) = -\frac{1}{2}\log \frac{\left(\int fg\right)^2}{\left(\int f^2\right)\left(\int g^2\right)}. \tag{4}$$

As we mentioned before, structure can be understood as regularities on the outcomes of a process. Therefore, the entropy related to the outcomes can be attributed in part to the underlying structure, and the rest to particular to each outcome as details or simple non related perturbations. Hence, we can think of the minimization of entropy as a means for finding such regularities. Suppose we are given a random variable S with PDF g, for which we want to find a description in terms of a PDF f with reduced entropy, that is, a variable X that captures the underlying structure of S. The principle of relevant information (PRI) formulates the above problem as a trade-off between the entropy $H_2(f)$ of X and its descriptive power about the observed random variable S in terms of their relative entropy $D_{\text{CS}}(f\|g)$. For a fixed PDF $g \in \mathcal{F}$ the objective is given by:

$$J(f) = H_2(f) + 2\lambda D_{\text{CS}}(f\|g), \tag{5}$$

where λ is the trade-off parameter. The minimization of J within a set of admissible PDFs $\mathcal{F}$ should lead to a function $f \in \mathcal{F}$ that has minimum entropy, but at the same time has maximum information gain about g. Nevertheless, as it is often the case, neither g nor a suitable space $\mathcal{F}$ are given directly. The only available information about g is encoded in a sample $S = \{x_i\}_{i=1}^N$, and some assumptions about the function class $\mathcal{F}$ must be made in order to obtain a tractable solution. The following section describes a solution to the problem that arises from the non-parametric estimator of the quantities in equation (3), which is based on weighted Parzen window method.

3 The Information Potential and Principle of Relevant Information

For the set $\mathcal{F}$ of probability density functions that are square integrable in $\mathbb{R}^n$, we can define the cross-information potential $\mathcal{V}$ (CIP), as a bilinear form that maps densities $f_i, f_j \in \mathcal{F}$ to the real numbers trough the integral,

$$\mathcal{V}(f_i, f_j) = \int_{\mathbb{R}^n} f_i(x) f_j(x) \mathrm{d}x \tag{6}$$

It is easy to see that for a basis of uniformly bounded, square integrable, probability density functions, $\mathcal{V}$ is a positive semidefinite function on the $\mathrm{span}\{\mathcal{F}\}$. Now consider the set $\mathcal{G} = \{g = \sum_{i=1}^m \alpha_i \kappa_\sigma(x_i, \cdot) | x_i \in \mathbb{R}^n,\ \sum_{i=1}^m \alpha_i = 1, \text{and } \alpha_i \geq 0\}$, where κ_σ is a "Parzen" type of kernel, that is κ_σ is symmetric, nonnegative, has bounded integral (can be normalized), and shift invariant with σ as the scale parameter. Clearly for any $g \in \mathcal{G}$ we have $\|g\|_2 \leq \|\kappa_\sigma(x, \cdot)\|_2$ thence $\mathcal{G}$ is bounded. However, if the $\mathcal{X}$ is non-compact our search space is also non compact.

The objective function for the principle of relevant information (5) can be written in terms of IP function. Using the Parzen based estimation, we restrict the search problem to $\mathcal{G} \subset \mathcal{F}$. In this case, we have that equation (5) can be rewritten as:

$$J(f) = -\log \mathcal{V}(f, f) - \lambda \log \frac{[\mathcal{V}(f, g)]^2}{\mathcal{V}(f, f)\mathcal{V}(g, g)} \tag{7}$$

straightforward manipulation of the terms yields an equivalent problem:

$$\arg\min_{f \in \mathcal{G}} [-(1 - \lambda) \log \mathcal{V}(f, f) - 2\lambda \log \mathcal{V}(f, g)] \tag{8}$$

Two important aspects of the above objective are: the choice of the kernel, shape and size σ, determines different scales for the analysis; the trade-off parameter λ defines a set of regimes for the possible solutions to the problem. As we previously mentioned, the only available information is contained in the sample $S = \{x_i\}_{i=1}^N$. An approximation of the target density g is then given by it weighted Parzen window estimator $\hat{g}(x) = \sum_{i=1}^N \alpha_i \kappa_\sigma(x_i, x)$, where where $\alpha_i \geq 0$ and $\sum_{i=1}^N \alpha_i = 1$, in our experiments we limit to $\alpha_i = 1/N$. To enforce compactness in our search space, we look for a solution f that has the same form of $\hat{g}$, that is

$$f(x) = \sum_{i=1}^N \beta_i f_i(x) = \sum_{i=1}^N \beta_i \kappa_\sigma(x_i, x). \tag{9}$$

where $\beta_i \geq 0$ and $\sum_{i=1}^N \beta_i = 1$. By fixing λ and evaluating the information potential between each pair $(x_i, x_j) \in S \times S$, we can rewrite (8) in matrix notation as:

$$\begin{aligned} &\min_{\boldsymbol{\beta}} \left[(\lambda - 1) \log \boldsymbol{\beta}^{\mathrm{T}} \mathbf{V} \boldsymbol{\beta} - 2\lambda \log \boldsymbol{\beta}^{\mathrm{T}} \mathbf{V} \boldsymbol{\alpha} \right] \\ &\text{subject to } \beta_i \geq 0 \text{ and } \sum_{i=1}^N \beta_i = 1 \end{aligned} \tag{10}$$

Notice that the form of the problem adopted in (10) is not a convex program, nevertheless it can be turned into an equivalent form that can be recognized as a convex program.

Proposition 3.1 *The convex program,*

$$\begin{aligned} \min_{\boldsymbol{\beta}} \ & \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} \\ \text{subject to } & \boldsymbol{\beta} \geq 0 \\ & \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} - \eta = 0 \\ & \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} - 1 = 0, \end{aligned} \tag{11}$$

is equivalent to (10)*, where* $\mathbf{q} = \mathbf{V}\boldsymbol{\alpha}$ *and some* $\eta > 0$*.*

Proof 3.1 *By definition* $\eta > 0$*, thus the constraint* $\log \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} = \log \eta$ *is equivalent to* $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} - \eta = 0$*. The positive semi definiteness of the information potential tell us that* $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} \geq 0$*, however, taking into account* $\mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} - \eta = 0$ *guarantees strict inequality; therefore the minimizers of* $\log \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}$ *and* $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}$ *on the constraint set defined in* (11) *are the same. Thence solving the following pseudo-convex program*

$$\begin{aligned} \min_{\boldsymbol{\beta}} \ & \log \boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta} \\ \text{subject to } & \boldsymbol{\beta} \geq 0 \\ & \log \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} = \log \eta \\ & \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} - 1 = 0, \end{aligned} \tag{12}$$

should yield the same solution. Now,The gradient of the objective in (10) *with respect to the weight vector* $\boldsymbol{\beta}$ *is,*

$$\nabla J(\boldsymbol{\beta}) = 2\left(\frac{\lambda - 1}{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\beta}}\right)\mathbf{V}\boldsymbol{\beta} - 2\left(\frac{\lambda}{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{V}\boldsymbol{\alpha}}\right)\mathbf{V}\boldsymbol{\alpha}. \tag{13}$$

By including the constraints $\mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} = 1$ *and* $\boldsymbol{\beta} \geq 0$*, for* $\lambda > 1$*, the set of KKT necessary conditions for local optimality in the Lagrangian* $L(\boldsymbol{\beta}, \boldsymbol{\mu}, \gamma) = J(\boldsymbol{\beta}) + \sum_{i=1}^{N} \mu_i c_i(\boldsymbol{\beta}) + \gamma e(\boldsymbol{\beta})$ *is*

$$\begin{cases} \frac{\partial}{\partial \boldsymbol{\beta}} L(\boldsymbol{\beta}, \boldsymbol{\mu}, \gamma) = \nabla J(\boldsymbol{\beta}) + \sum_{i=1}^{N} \mu_i \frac{\partial}{\partial \boldsymbol{\beta}} c_i(\boldsymbol{\beta}) + \gamma \frac{\partial}{\partial \boldsymbol{\beta}} e(\boldsymbol{\beta}) = 0, \\ \frac{\partial}{\partial \boldsymbol{\mu}} L(\boldsymbol{\beta}, \boldsymbol{\mu}, \gamma) = \mathbf{c}(\boldsymbol{\beta}) \leq 0, \\ \boldsymbol{\mu}^{\mathrm{T}}\mathbf{c}(\boldsymbol{\beta}) = 0 = -\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta}, \\ \boldsymbol{\mu} \geq 0, \\ \frac{\partial}{\partial \gamma} L(\boldsymbol{\beta}, \boldsymbol{\mu}, \gamma) = e(\boldsymbol{\beta}) = 0 = \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} - 1. \end{cases} \tag{14}$$

There are two possible cases for each β_i^*

- $\beta_i^* > 0$.
  For which $\mu_i^* = 0$ *and*

$$2\frac{t_i}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*} - 2\left(\frac{\lambda}{\lambda - 1}\right)\frac{q_i}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{q}} + \gamma = 0, \tag{15}$$

  where $\mathbf{t} = \mathbf{V}\boldsymbol{\beta}$*.*

- $\beta_i^* = 0$.
  Yields

$$2\frac{t_i}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*} - 2\left(\frac{\lambda}{\lambda - 1}\right)\frac{q_i}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{q}} - \mu_i^* + \gamma = 0. \tag{16}$$

Notice that $\gamma = 2\left(\frac{1}{\lambda-1}\right)$*, therefore,*

$$2\frac{\mathbf{V}\boldsymbol{\beta}^*}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*} - 2\left(\frac{\lambda}{\lambda-1}\right)\frac{\mathbf{q}}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{q}} - \boldsymbol{\mu}^* + 2\frac{\mathbf{1}}{\lambda-1} = 0. \tag{17}$$

Pre-multiplying (17) *by* $(\boldsymbol{\beta} - \boldsymbol{\beta}^*)$ *in the constraint set, yields the following set of conditions*

$$\begin{aligned}
(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^{\mathrm{T}}\left[2\frac{\mathbf{V}\boldsymbol{\beta}^*}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*} - \boldsymbol{\mu}^*\right] &\geq 0,\ \forall\boldsymbol{\beta} \geq 0:\ \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta} = \eta, \mathbf{1}^{\mathrm{T}}\boldsymbol{\beta} = 1 \\
\boldsymbol{\mu}^{*\mathrm{T}}\boldsymbol{\beta}^* &= 0 \\
-\boldsymbol{\beta}^* &\leq 0 \\
\boldsymbol{\mu}^* &\geq 0 \\
\log \mathbf{q}^{\mathrm{T}}\boldsymbol{\beta}^* &= \log \eta \\
\mathbf{1}^{\mathrm{T}}\boldsymbol{\beta}^* &= 1,
\end{aligned} \tag{18}$$

which by Theorem A.1 are sufficient conditions for the solution of a pseudo-convex function defined on an open set with convex inequality constraints, that in our case corresponds to (12)

□

Two important results come from the above proposition. One is obvious from the statement in the proposition that tells us there exist an equivalent convex program that solves (10). But even better is the one that comes as a byproduct of the proof. The KKT first order conditions in (14) are necessary and sufficient to solve (10).

4 Decomposition into Smaller Subproblems

In the proof of Proposition 3.1 we solve a more convenient form of (10), for which we factorize $(\lambda-1)$ from the objective. If we derive the the solution for the original problem the two cases (15) and (16) are replaced by:

- $\beta_i^* > 0$.
  With $\mu_i^* = 0$ and

  $$2\frac{\lambda-1}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*}t_i - 2\frac{\lambda}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{q}}q_i + \gamma = 0, \tag{19}$$

  where $\mathbf{t} = \mathbf{V}\boldsymbol{\beta}$.
- $\beta_i^* = 0$.

  $$2\frac{\lambda-1}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*}t_i - 2\frac{\lambda}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{q}}q_i - \mu_i^* + \gamma = 0. \tag{20}$$

  Note that combining (19) and (20) with the optimal $\boldsymbol{\beta}^*$ we have that $\gamma = 2$, using this fact along with the non-negativity of $\boldsymbol{\mu}$, the following condition should hold,

  $$At_i - Bq_i > 1, \tag{21}$$

  where $A = \frac{\lambda-1}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{V}\boldsymbol{\beta}^*}$ and $B = \frac{\lambda}{\boldsymbol{\beta}^{*\mathrm{T}}\mathbf{q}}$.

Let's partition the set of indexes of the entries of $\boldsymbol{\beta}$ into W, the working set, and P the complementary set of inactive elements. Then $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_W^{\mathrm{T}},\ \boldsymbol{\beta}_P^{\mathrm{T}}\right)^{\mathrm{T}}$, for which we define the following subproblem:

$$\begin{aligned} \min_{\boldsymbol{\beta}_W} \Big[& (\lambda - 1) \log \left(\boldsymbol{\beta}_W^{\mathrm{T}} \mathbf{V}_{WW} \boldsymbol{\beta}_W + 2 \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{V}_{PW} \boldsymbol{\beta}_W + \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{V}_{PP} \boldsymbol{\beta}_P \right) + \\ & - 2\lambda \log \left(\boldsymbol{\beta}_W^{\mathrm{T}} \mathbf{q}_W + \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{q}_P \right) \Big] \\ \text{subject to} \quad & - \boldsymbol{\beta}_W \leq 0, \quad \text{and} \quad \left[\boldsymbol{\beta}_W^{\mathrm{T}} \mathbf{1} + \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{1} \right] = 1 \end{aligned} \tag{22}$$

Similar remarks to the ones made in [6] can be obtained for (22):

- The terms $\varphi_A = \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{V}_{PP} \boldsymbol{\beta}_P$ and $\varphi_B = \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{q}_P$ are constant in the subproblem
- The computation of $2\boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{V}_{PW} \boldsymbol{\beta}_W$ is independent of the size of P and also of the number of nonzero β_i's
- Replacing β_i, with $i \in W$ with β_j with $j \in P$ leaves the cost unchangend and the feasibility remains intact.
- If the subproblem is optimal before the above replacement, the new subproblem is optimal if and only if β_j satisfies the optimality conditions.

The so called "*Buld down*" step is rather obvious. Now the "*Buld up*" step that states that moving a variable from P to W gives an strict improvement in the cost when the subproblem is re-optimized. In our case we can justify the build up since we prove that the KKT first order conditions are necessary and sufficient for a solution to (10). These build down and build up steps suggest an strategy for optimizing (10) by solving smaller subproblems. At each iteration, solve a subproblem that include a constraint violator picked from the complementary set P. Iterate until optimality conditions are satisfied up two some desired level of accuracy.

5 Sequential Minimal Optimization

In the previous section (4), the optimization problem related to the principle of information, was decomposed into smaller subproblems that can be solved iteratively to achieve the solution to the full problem. An important characteristic of such decomposition is that the size of the working set W and the complementary set P, are independent of the number of support vectors in the solution, that is, the β_i^*'s greater than zero. The sequential minimal optimization proposed in [7] chooses the smallest subproblem that can be solved at each iteration. This corresponds to solving for two variables at the time, which can be found analytically. The latter is of particular appeal to the solve PRI since our cost does not have an standard form as it is the case for SVMs (quadratic program), therefore, we cannot resort to off the shelf solvers for our problem.

Without loss of generality we will refer to our variables in the working set as β_1 and β_2 and the complementary set as P. By the equality constraint in the subproblem (22) we have that $\beta_1 + \beta_2 = 1 - \boldsymbol{\beta}_P^{\mathrm{T}} \mathbf{1} = w$ and thence $\beta_1 = w - \beta_2$.

Let us denote by $\bar{\beta}_i$ the value of β_i from the previous iteration. We can formulate the subproblem in terms of β_2 as:

$$\begin{aligned} &\min_{\beta_2} \left[(\lambda-1)\log A(\beta_2) - 2\lambda \log B(\beta_2)\right] \\ &\text{subject to} \quad 0 \leq \beta_2 \leq w, \quad \text{and} \quad w = \bar{\beta}_1 + \bar{\beta}_2, \end{aligned} \tag{23}$$

with

$$A(\beta_2) = \beta_2^2(V_{11}+V_{22}-2V_{12})+2\beta_2(w(V_{12}-V_{11})+(\upsilon_2-\upsilon_1))+w^2V_{11}+2w\upsilon_1+\varphi_A$$

where $\upsilon_i = \mathbf{V}_i\bar{\boldsymbol{\beta}} - V_{1i}\bar{\beta}_1 - V_{1i}\bar{\beta}_2$, and $\varphi_A = {\boldsymbol{\beta}_P}^{\mathrm{T}}\mathbf{V}_{PP}\boldsymbol{\beta}_P$; and

$$B(\beta_2) = \beta_2(q_2 - q_1) + wq_1 + \varphi_B$$

where $\varphi_B = {\boldsymbol{\beta}_P}^{\mathrm{T}}\mathbf{q}_P$. The solution to problem (23) lies on the line segment $\beta_1 = w - \beta_2$ with $0 \leq \beta_2 \leq w$. Computing the derivative of the objective in (23) yields a second order polynomial on β_2 (Details are given in Appendix B), thus solving

$$c_2\beta_2^2 + c_1\beta_2 + c_0 = 0 \tag{24}$$

with coefficients:

$$\begin{aligned} c_2 &= -2(V_{11}+V_{22}-2V_{12})(q_2-q_1) \\ c_1 &= 2(\lambda-1)(wq_1+\varphi_B)(V_{11}+V_{22}-2V_{12}) + \\ &\quad -2(\lambda+1)(w(V_{12}-V_{11})+(\upsilon_2-\upsilon_1))(q_2-q_1) \\ c_0 &= 2(\lambda-1)(w(V_{12}-V_{11})+(\upsilon_2-\upsilon_1))(wq_1+\varphi_B) + \\ &\quad -2\lambda(w^2V_{11}+2w\upsilon_1+\varphi_A)(q_2-q_1) \end{aligned}$$

conveys candidate solutions that ought be checked along with the end points of the line segment. Let r_1 and r_2 be the roots of (24). Ruling out cases with complex numbers, we have:

$$L = \min\{r_1,\ r_2\} \quad \text{and} \quad U = \max\{r_1,\ r_2\}$$

the candidate solutions are,

$$s_1 = \begin{cases} 0 & L \leq 0 \\ L & 0 < L < w \\ w & L \geq w \end{cases} \quad \text{and} \quad s_2 = \begin{cases} 0 & U \leq 0 \\ U & 0 < U < w \\ w & U \geq w \end{cases} \tag{25}$$

If $s_1 \neq s_2$ we check $J(s_i) = [(\lambda-1)\log A(s_i) - 2\lambda\log B(s_i)]$ and

$$\beta_2 = \arg\min_{s_i \in \{s_1, s_2\}} \{J(s_i)\} \tag{26}$$

otherwise, $\beta_2 = s_1 = s_2$.

5.1 SMO algorithm

The algorithm can be described into three basic steps:

Step 1: Initialization

$$\begin{aligned}
\mathbf{q} &\leftarrow \mathbf{V}\boldsymbol{\alpha} \\
\mathbf{f} &\leftarrow \mathbf{q} \\
\boldsymbol{\beta} &\leftarrow \boldsymbol{\alpha} \\
IP(\boldsymbol{\beta}) &\leftarrow \boldsymbol{\beta}^{\mathrm{T}}\mathbf{q} \\
CIP(\boldsymbol{\beta}) &\leftarrow IP(\boldsymbol{\beta})
\end{aligned}$$

Step 2: Constants within an iteration

$$\begin{aligned}
\upsilon_i &\leftarrow f_i - V_{1i}\bar{\beta}_1 - V_{2i}\bar{\beta}_2 \\
\varphi_A &\leftarrow IP(\bar{\boldsymbol{\beta}}) - (2(\bar{\beta}_2 f_1 + \bar{\beta}_2 f_2) - (\bar{\beta}_1^2 V_{11} + \bar{\beta}_2^2 V_{22} + 2\bar{\beta}_1 V_{12}\bar{\beta}_2)) \\
\varphi_B &\leftarrow CIP(\bar{\boldsymbol{\beta}}) - (\bar{\beta}_1 q_1 + \bar{\beta}_2 q_2) \\
w &\leftarrow \bar{\beta}_1 + \bar{\beta}_2
\end{aligned}$$

Step 3: Updates

$$\begin{aligned}
\beta_2 &\leftarrow \text{solution described in (26)} \\
\beta_1 &\leftarrow w - \beta_2 \\
\mathbf{f} &\leftarrow \bar{\mathbf{f}} + (\bar{\beta}_2 - \beta_2)\mathbf{V}_1^{\mathrm{T}} + (\beta_2 - \bar{\beta}_2)\mathbf{V}_2^{\mathrm{T}} \\
IP(\boldsymbol{\beta}) &\leftarrow \varphi_A + (2(\beta_1 f_1 + \beta_2 f_2) - (\beta_1^2 V_{11} + \beta_2^2 V_{22} + 2\beta_1 V_{12}\beta_2)) \\
CIP(\boldsymbol{\beta}) &\leftarrow \varphi_B + (\beta_1 q_1 + \beta_2 q_2)
\end{aligned}$$

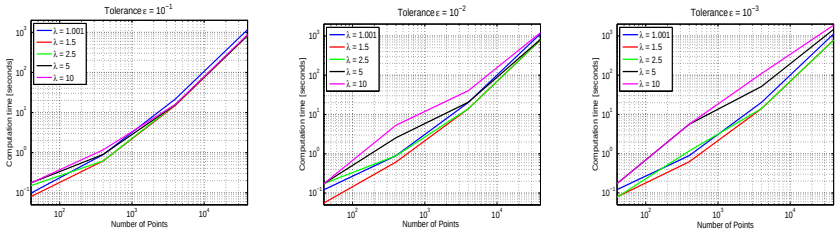
Steps 2 and 3 are iterated for different working sets chosen according to some heuristics that are described below.

5.2 Selecting the working set

There are two type of constraint violations, an equality constraint (19) if $\beta_i > 0$, and the inequality constraint (21) if $\beta_i = 0$. The constraint violations are easy to compute at each iteration. Let $\boldsymbol{\xi}$ be defined as

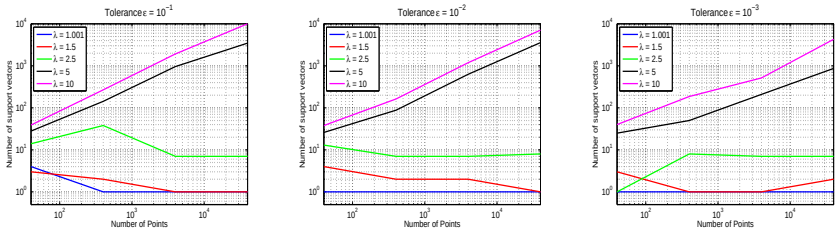
$$\boldsymbol{\xi} = 2\frac{\lambda - 1}{IP(\boldsymbol{\beta})}\mathbf{f} - 2\frac{\lambda}{CIP(\boldsymbol{\beta})}\mathbf{q}, \tag{27}$$

the constraint qualifications are $\xi_i = 2$ if $\beta_i > 0$, and $\xi_i > 2$ if $\beta_i = 0$. In teh description of our algorithm we chose to initialize $\boldsymbol{\beta}$ with the same values of $\boldsymbol{\alpha}$. However our cost function suggest that points for which q_i is large will be expected to become support vectors, that is $\beta_i > 0$. We can then use $\boldsymbol{\beta} = \mathbf{q}/(\mathbf{q}^{\mathrm{T}}\mathbf{1})$ as the initial guess. However this would imply the computation of $\mathbf{f}$ at the initialization. It is customary to choose $\boldsymbol{\alpha} = \frac{1}{N}\mathbf{1}$. Then at the initial iteration all constraints will be violated (unless $\lambda \to \infty$). One pass through the whole set taking pairs of indexes $(i,\ j)$, where i correspond to a descending order of the samples according to $\mathbf{f}$ and j's taken at random will create the first stage of sparseness in our weight vector $\boldsymbol{\beta}$; this is our first heuristic. After this pass, we can check whether (21) is satisfied for the current β_i's that are zero. A second stage suggest checking the within the set of samples with $\beta_i > 0$, and for which $\xi_i > 2$ since they are most likely to vanish. We will stop when conditions are fulfilled within ϵ tolerance.



(a) (b) (c)

Figure 1: Computation times for different tolerance levels and sample sizes



(a) (b) (c)

Figure 2: Number of support vectors for different tolerance levels and sample sizes

6 Experiments

6.1 Synthetic Data

Here, we are concerned with the computation of the principle of relevant information on large sample sizes. The purpose of this experimental setup is to observe the behavior of the algorithm in terms of λ which controls the number of nonzero weights and therefore the number of equality constraints that are much harder to satisfy. Data is obtained by sampling from a two dimensional mixture of three Gaussians with centers $(0,\ 0)$, $(3,\ 3)$, and $(-6,\ 4)$; spherical covariances $0.8^2\mathbf{I}$, $1.2^2\mathbf{I}$, and $\mathbf{I}$; and mixing proportions 0.2, 0.3, and 0.5, respectively. The kernel employed in our experiments is the Gaussian kernel $\kappa(x, y) = \exp\left(-\frac{1}{2\sigma^2}\|x - y\|^2\right)$, with $\sigma = 0.2$. Figure 1 depicts the computation times for different tolerance levels on the constraint violations as well as various sample sizes and trade off parameter λ. Figure 2 shows the final number of support vectors (nonzero weights) when the above mentioned parameters are varied.

First notice that the kernel size σ was kept fixed regardless of the size of the sample. this allows for studying the algorithm behavior in terms of sparsity of data, which in this case correspond to small sample sizes. The tolerance level has a clear effect on the computation time, but more interesting is the effect on the number of support vectors which reduces when the level of accuracy increase. On the small sample regime, the increment on the computation time due to the more demanding tolerance level $\epsilon = 10^{-3}$ can be attributed to the scarcity of data which makes the cost function much more sensitive to any change in the weight vector $\boldsymbol{\beta}$. In terms of computational complexity the algorithm behaves within the reasonable levels, In

the experiments carried we bound the maximum number of iterations by $N \log N$. This bound upper bound was never attained by the larger sample sizes and only reached by small sample sizes on the most demanding scenarios, that is, small ϵ and large λ, since the trade off parameter λ is closely related with the number of support vectors and thus the proportion of constraint violators increases.

7 Conclusions

We have introduced a sequential minimal optimization algorithm for the principle of relevant information based on weighted density estimation. In order to guarantee convergence of the algorithm, we show that the Karush-Kuhn-Tucker first order optimality condition are necessary and sufficient in our formulation. In proving this, we found there exist a convex program that yields the same solution, however this result is not yet applicable in an alternative implementation. Results show that computational complexity is manageable even for sample sizes of several tens of thousands. The very important feature is that elements of the Gram matrix are computed at request and do not need to be stored, nevertheless, speed improvements can be achieved by using a cache that temporarily stores frequently visited samples. Several improvements in terms of speed by better selection heuristics and memory trade offs can be pursued in future implementations.

References

[1] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[2] S. Fine and K. Scheinberg. Efficient svm training using low-rank kernel representations. *JMLR*, 2:243–264, 2001.

[3] T. Joachims. *Advances in Kernel Methods: Support Vector Learning*, chapter 11: Making Large-Scale SVM Learning Practical. The MIT Press, 1999.

[4] E. Lutwak, D. Yang, and G. Zhang. Cramér-rao and moment-entropy inequalities for renyi entropy and generalized fisher information. *IEEE Transactions on Information Theory*, 51(2):473–478, February 2005.

[5] O. Mangasarian. *Nonlinear Programming*. Systems and Science. McGraw-Hill, 1969.

[6] E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *IEEE Workshop on neural Networks for Signal Processing*, pages 276–285, 1997.

[7] J. C. Platt. Sequential minimal optimization: A fast algorithm for training support vector machines. Technical report, Microsoft Research, April 1998.

[8] S. M. Rao. *Unsupervised Learning: An Information Theoretic Framework*. PhD thesis, University of Florida, 2008.

[9] A. Rényi. On measures of entropy and information. In *Proceedings of the fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 547–561, Berkeley, 1961. University of California Press.

[10] S. Seth and J. Principe. On speeding up computation in information theoretic learning. In *IJCNN*, 2009.

[11] A. J. Smola and B. Schölkopf. Sparse greedy matrix approximation for machine learning. *ICML*, pages 911–918, 2000.

[12] C. K. Williams and M. Seeger. Using the nyström method to speed up kernel machines. In *NIPS*, pages 682–688, 2000.

A Sufficient Conditions for Pseudo-Convex Programs

The following theorem is extracted from [5] Chapter 10.

Theorem A.1 *Let $\mathcal{X}^0$ be an open set in $\mathbb{R}^n$ an let f and g be respectively and scalar and a m-dimensional vector function both defined in $\mathcal{X}^0$. Let $x^* \in \mathcal{X}^0$, $I = \{i|\ g_i(x^*) = 0\}$, f be pseudo-convex at x^*, and g be differentiable and quasi-convex at x^*. If there exists a $\boldsymbol{\mu}^* \in \mathbb{R}^m$ such that the pair $(x^*, \boldsymbol{\mu}^*)$ satisfy the following conditions:*

$$\begin{aligned} \left[\nabla f(x^*) + \boldsymbol{\mu}^{*\mathrm{T}} \mathrm{D}g(x^*)\right](x - x^*) &\geq 0,\ \forall x \in \mathcal{X}^0;\ g(x) \leq 0 \quad (28)\\ \boldsymbol{\mu}^{*\mathrm{T}} g(x^*) &= 0\\ g(x^*) &\leq 0\\ \boldsymbol{\mu}^* &\geq 0 \end{aligned}$$

Then, x^* is a solution of the following minimization problem

$$\begin{aligned} &\min_{x \in \mathcal{X}^0}\ f(x)\\ &\text{subject to}\ g(x) \leq 0. \end{aligned} \quad (29)$$

Proof A.1 *Let $I = \{i|\ g_i(x^*) = 0\}$, $J = \{j|\ g_i(x^*) < 0\}$, thence $I \cup J = \{1, \ldots, m\}$ since $\boldsymbol{\mu}^* \geq 0$, $g(x^*) \leq 0$, and $\boldsymbol{\mu}^* \geq 0$, we have that $\{\mu_j\}_{j \in J} = 0$, and from quasi convexity of g at x^*, the gradients of g_i at x^* for $i \in I$ are orthogonal to tangent planes to the level sets defined by $g_i(x) = 0$ and therefore for any feasible point $x \in \mathcal{X}^0$ and $g(x) \leq 0$, $\mathrm{D}g_I(x^*)(x - x^*) \leq 0$, by non-negativity of $\boldsymbol{\mu}$ and since $\boldsymbol{\mu}_J = 0$ we have:*

$$\begin{aligned} \boldsymbol{\mu}_I^{*\mathrm{T}} \mathrm{D}g_I(x^*)(x - x^*) &\leq 0 \quad (30)\\ \boldsymbol{\mu}_J^{*\mathrm{T}} \mathrm{D}g_J(x^*)(x - x^*) &= 0\\ \boldsymbol{\mu}^{*\mathrm{T}} \mathrm{D}g(x^*)(x - x^*) &= \left[\boldsymbol{\mu}_I^{*\mathrm{T}} \mathrm{D}g_I(x^*) + \boldsymbol{\mu}_J^{*\mathrm{T}} \mathrm{D}g_J(x^*)\right](x - x^*) \leq 0. \end{aligned}$$

Finally, since $\left[\nabla f(x^*) + \boldsymbol{\mu}^{*\mathrm{T}} \mathrm{D}g(x^*)\right](x - x^*) \geq 0$ for all $x \in \mathcal{X}^0$ and $g(x) \leq 0$, we need that $\nabla f(x^*)(x - x^*) \geq 0$ and thus by pseudo-convexity of f implying that $f(x) \geq f(x^*)$ for all $x \in \mathcal{X}^0$ such that $g(x) \leq 0$.

□

A generalization of the Kuhn-Tucker sufficient optimality criterion follows from the above theorem by replacing condition (28) with

$$\nabla f(x^*) + \boldsymbol{\mu}^{*\mathrm{T}} \mathrm{D}g(x^*) = 0 \quad (31)$$

B Details of the Solution to the Minimal Subproblem

We refer to the objective in (23) as,

$$J(\beta_2) = (\lambda - 1)\log A(\beta_2) - 2\lambda \log B(\beta_2). \tag{32}$$

Taking the derivative of $J(\beta_2)$ and equating to zero yields:

$$\frac{\mathrm{d}}{\mathrm{d}\beta_2} J(\beta_2) = 0 = (\lambda - 1)B(\beta_2)\frac{\mathrm{d}}{\mathrm{d}\beta_2} A(\beta_2) - 2\lambda A(\beta_2)\frac{\mathrm{d}}{\mathrm{d}\beta_2} B(\beta_2) \tag{33}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}\beta_2} A(\beta_2) = 2(\beta_2(V_{11} + V_{22} - V_{12}) + (w(V_{12} - V_{11}) + (\upsilon_2 - \upsilon_1))) \tag{34}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\beta_2} B(\beta_2) = q_2 - q_1 \tag{35}$$

Expanding and rearranging yields (24)