Proof of Convergence for Kernelized M3L

Bharath Hariharan   Lihi Zelnik-Manor   S.V.N Vishwanathan   Manik Varma

July 20, 2010

We now describe the convergence of the algorithm for kernelized M3L [2]. It closely follows the proof of convergence of SMO [3].

1 Notation

We denote vectors in bold small letters, for example $\mathbf{v}$. If $\mathbf{v}$ is a vector of dimension d, then $v_k, k \in \{1, \ldots, d\}$ is the kth component of $\mathbf{v}$, and $\mathbf{v}_I, I \subseteq \{1, \ldots, d\}$ denotes the vector with components $v_k, k \in I$ (with the v_k's arranged in the same order as in $\mathbf{v}$). Similarly, matrices will be written in bold capital letters, for example $\mathbf{A}$. If $\mathbf{A}$ is an $m \times n$ matrix, then A_{ij} represents the ijth entry of $\mathbf{A}$, and $\mathbf{A}_{IJ}$ represents the matrix with entries $A_{ij}, i \in I, j \in J$.

A sequence is denoted as $\{a^n\}$, and a^n is the nth element of this sequence. If $\hat{a}$ is a limit point of the sequence, we write $a^n \to \hat{a}$.

2 The Optimization Problem

The dual that we are trying to solve is:

$$\max_{\boldsymbol{\alpha}} 2\sum_{l=1}^{L} \boldsymbol{\alpha}_l^t \mathbf{1} - 2\sum_{l=1}^{L}\sum_{k=1}^{L} R_{lk}\boldsymbol{\alpha}_l^t \mathbf{Y}_l \mathbf{K} \mathbf{Y}_k \boldsymbol{\alpha}_k \tag{1}$$

s.t

$$0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}$$

where $\boldsymbol{\alpha}_l = [\alpha_{1l}, \ldots \alpha_{Nl}]$, $\mathbf{Y}_l = diag([y_{1l} \ldots y_{Nl}])$ and $\mathbf{K} = \phi(\mathbf{X})^t \phi(\mathbf{X})$. Making the substitution $\theta_{il} = 2y_{il}\alpha_{il}$, we get the following optimization problem:

$$\max_{\boldsymbol{\theta}} \sum_{i=1}^{N}\sum_{l=1}^{L} \theta_{il} y_{il} - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{L}\sum_{l=1}^{L} \theta_{ik}\theta_{jl} R_{kl} K_{ij} \tag{2}$$

s.t

$$0 \leq \theta_{ik} \leq 2C \qquad \forall i, k \text{ s.t } y_{ik} > 0$$

$$-2C \leq \theta_{ik} \leq 0 \qquad \forall i, k \text{ s.t } y_{ik} < 0$$

This can be written as the following optimization problem:

Problem:

$$\max_{\boldsymbol{\theta}} f(\boldsymbol{\theta}) = -\frac{1}{2}\boldsymbol{\theta}^t \mathbf{Q} \boldsymbol{\theta} + \mathbf{y}^t \boldsymbol{\theta} \tag{3}$$

s.t

$$\mathbf{l} \leq \boldsymbol{\theta} \leq \mathbf{u}$$

Here the vector $\boldsymbol{\theta} = [\theta_{11} \ldots \theta_{1L}, \theta_{21}, \ldots \theta_{NL}]^t$, $\mathbf{y} = [\mathbf{y}_1^t \ldots \mathbf{y}_N^t]^t$ and $\mathbf{Q} = \mathbf{K} \otimes \mathbf{R}$ where $\otimes$ is the Kronecker product. $\mathbf{l}$ and $\mathbf{u}$ are NL- dimensional vectors with entries:

$$l_{ik} = \begin{cases} 0 & \text{if } y_{ik} > 0 \\ -2C & \text{if } y_{ik} < 0 \end{cases} \tag{4}$$

$$u_{ik} = \begin{cases} 2C & \text{if } y_{ik} > 0 \\ 0 & \text{if } y_{ik} < 0 \end{cases} \tag{5}$$

3 The algorithm

We will prove the convergence result for the algorithm given in Algorithm 1. We start with $\boldsymbol{\theta}^1 = \mathbf{0}$, and generate a sequence of vectors $\boldsymbol{\theta}^n$, where $\boldsymbol{\theta}^n$ is the vector after the $(n-1)$th iteration. The gradient $\nabla f(\boldsymbol{\theta}^n)$, denoted by $\mathbf{g}^n$, is stored. The projected gradient $\nabla^P f(\boldsymbol{\theta}^n)$, denoted by $\tilde{\mathbf{g}}^n$, is computed from the gradient on the fly. $\tau > 0$ is a user-defined parameter that determines when we stop. The algorithm terminates when all projected gradients are less than τ.

We assume that $\mathbf{R}$ and $\mathbf{K}$ are both positive definite matrices. The eigenvalues of $\mathbf{Q}$ are then $\lambda_i\mu_j$, where λ_i are the eigenvalues of $\mathbf{K}$ and μ_j are the eigenvalues of $\mathbf{R}$. Because all eigenvalues of both $\mathbf{R}$ and $\mathbf{K}$ are positive, so are the eigenvalues of $\mathbf{Q}$ and thus $\mathbf{Q}$ is positive definite. Therefore, using Lemma 8 (refer to Appendix A), we have that $\boldsymbol{\theta}$ is optimal iff $\boldsymbol{\theta}$ is feasible and $\nabla^P f(\boldsymbol{\theta}) = \mathbf{0}$.

Our algorithm will not reach an exact solution, and therefore the KKT conditions will not be satisfied exactly. Call a (feasible) solution $\boldsymbol{\theta}$ τ-optimal if the KKT conditions are satisfied to a precision of τ, i.e, $\mid \nabla^P_{ik} f(\boldsymbol{\theta}) \mid < \tau$ for each i, k. Then our algorithm terminates when it reaches a τ-optimal solution. Lemma 10 in Appendix A shows that by making τ arbitrarily close to 0, we can make our solution arbitrarily close to the optimum. In the following sections we will show that our algorithm will reach the τ-optimal solution in a finite number of steps.

4 Convergence

In this section we prove that the sequence of vectors $\boldsymbol{\theta}^n$ converges.

Algorithm 1 Our algorithm

```
θ_ik ← 0        ∀i, k
repeat
  Pick i, k such that | g̃_ik | ≥ τ
  if ∃ j such that K_ij ≠ sqrt(K_ii K_jj) and g̃_jk ≠ 0 then
    Pick one such j.
    Optimize w.r.t θ_ik and θ_jk
  else
    Optimize w.r.t θ_ik
  end if.
  Update g_ik      ∀i, k
until | g̃_ik | < τ      ∀i, k
```

Note the following:

- In each iteration of the algorithm, we optimize over a set of variables, which may either be a single variable θ_{ik} or a pair of variables $\{\theta_{ik}, \theta_{jl}\}$.
- The projected gradient of all the chosen variables is non zero at the start of the iteration.
- At least one of the chosen variables has projected gradient with magnitude greater than τ.

Consider the nth iteration. Denote by B the set of indices of the variables chosen: $B = \{(i,k)\}$ or $B = \{(i,k),(j,k)\}$. Without loss of generality, reorder variables so that the variables in B occupy the first few indices. In the nth iteration, we optimize f over the variables in B keeping the rest of the variables constant. Thus we have to maximize $h(\boldsymbol{\Delta}_B) = f(\boldsymbol{\theta}^n + [\boldsymbol{\Delta}_B^t,\ 0]^t) - f(\boldsymbol{\theta}^n)$. This amounts to solving the optimization problem:

$$\max_{\boldsymbol{\Delta}_B} h(\boldsymbol{\Delta}_B) = -\frac{1}{2}\boldsymbol{\Delta}_B^t \mathbf{Q}_{BB}\boldsymbol{\Delta}_B - \boldsymbol{\Delta}_B^t(\mathbf{Q}\boldsymbol{\theta}^n)_B + \mathbf{y}_B^t\boldsymbol{\Delta}_B \quad (6)$$

s.t

$$\mathbf{l}_B - \boldsymbol{\theta}_B \leq \boldsymbol{\Delta}_B \leq \mathbf{u}_B - \boldsymbol{\theta}_B$$

Assuming that $R_{kk} > 0\ \forall k$ and $K_{ii} > 0\ \forall i$, $\mathbf{Q}_{BB}$ is positive definite. This can be seen as follows. B has either one or two elements. In the first case, $B = \{(i,k)\}$ and $\mathbf{Q}_{BB} = R_{kk}K_{ii} > 0$ and hence $\mathbf{Q}_{BB}$ is positive definite. In the second case, suppose $B = \{(i,k),(j,k)\}$. The matrix $\mathbf{Q}_{BB}$ is given by $\begin{bmatrix} K_{ii}R_{kk} & K_{ij}R_{kk} \\ K_{ij}R_{kk} & K_{jj}R_{kk} \end{bmatrix}$. Because $K_{ij}^2 \neq K_{ii}K_{ij}$ (refer Algorithm 1) we have that $\mathbf{Q}_{BB}$ must be positive definite.

Hence by Lemma 8 in Appendix A, $\boldsymbol{\Delta}_B^*$ optimizes (6) iff it is feasible and

$$\nabla^P h(\boldsymbol{\Delta}_B^*) = \mathbf{0} \quad (7)$$

Then we have that $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n + \boldsymbol{\Delta}^*$, where $\boldsymbol{\Delta}^* = [\boldsymbol{\Delta}_B^{*t},\ \mathbf{0}^t]^t$. Now note that since $\mathbf{g}_B^n = -(\mathbf{Q}\boldsymbol{\theta}^n)_B + \mathbf{y}_B$

$$h(\boldsymbol{\Delta}_B) = -\frac{1}{2}\boldsymbol{\Delta}_B^t\mathbf{Q}_{BB}\boldsymbol{\Delta}_B + \boldsymbol{\Delta}_B^t\mathbf{g}_B^n \quad (8)$$

$$\Rightarrow \nabla h(\boldsymbol{\Delta}_B) = -\mathbf{Q}_{BB}\boldsymbol{\Delta}_B + \mathbf{g}_B^n \quad (9)$$

Also,

$$\begin{aligned}\mathbf{g}_B^{n+1} &= -(\mathbf{Q}\boldsymbol{\theta}^{n+1})_B + \mathbf{y}_B \quad (10)\\ &= -(\mathbf{Q}(\boldsymbol{\theta}^n + [\boldsymbol{\Delta}_B^{*t},\ \mathbf{0}^t]^t))_B + \mathbf{y}_B\end{aligned}$$

$$\Rightarrow \mathbf{g}_B^{n+1} = \mathbf{g}_B^n - \mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* = \nabla h(\boldsymbol{\Delta}_B^*) \quad (11)$$

Then (11) means that:

$$\tilde{\mathbf{g}}_B^{n+1} = \nabla^P h(\boldsymbol{\Delta}_B^*) \quad (12)$$

Using (7)

$$\tilde{\mathbf{g}}_B^{n+1} = \nabla^P h(\boldsymbol{\Delta}_B^*) = \mathbf{0} \quad (13)$$

This leads us to the following lemma:

Lemma 1. *Let $\boldsymbol{\theta}^n$ be the solution at the start of the nth iteration. Let B be the set of indices of the variables over which we optimize. Let the updated solution be $\boldsymbol{\theta}^{n+1}$. Then*

1. $\tilde{\mathbf{g}}_B^{n+1} = \mathbf{0}$
2. $\boldsymbol{\theta}^{n+1} \neq \boldsymbol{\theta}^n$
3. *If* $l_{jk} < \theta_{jk}^{n+1} < u_{jk}$ *then* $g_{jk}^{n+1} = 0\ \forall (j,k) \in B$

Proof. 1. This follows directly from (13).

2. If $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n$, then $\boldsymbol{\Delta}_B^* = \mathbf{0}$ and so, from (11), $\mathbf{g}_B^{n+1} = \nabla h(\mathbf{0}) = \mathbf{g}_B^n$. This means that from (13) $\tilde{\mathbf{g}}_B^n = \tilde{\mathbf{g}}_B^{n+1} = \mathbf{0}$. But this is a contradiction since we required that all variables in the chosen set have non zero projected gradient before the start of the iteration.

3. Since the final projected gradients are 0 for all variables in the chosen set (from (13)), if $l_{jk} < \theta_{jk}^{n+1} < u_{jk}$ then $g_{jk}^{n+1} = 0\ \forall (j,k) \in B$ □

Lemma 2. *In the same setup as the previous lemma, $f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) \geq \sigma\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n\|^2$, for some fixed $\sigma > 0$.*

Proof.

$$f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) = h(\boldsymbol{\Delta}_B^*) = -\frac{1}{2}\boldsymbol{\Delta}_B^{*t}\mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* + \boldsymbol{\Delta}_B^{*t}\mathbf{g}_B^n \quad (14)$$

where $\boldsymbol{\Delta}_B^*$ is the optimum solution of Problem (6). Now, note that since $\boldsymbol{\Delta}_B^*$ is feasible and $\mathbf{0}$ is feasible, we have, using Lemma 9 in Appendix A,

$$\boldsymbol{\Delta}_B^{*t}\nabla h(\boldsymbol{\Delta}_B^*) \geq 0 \quad (15)$$

$$\Rightarrow -\boldsymbol{\Delta}_B^{*t}\mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* + \boldsymbol{\Delta}_B^{*t}\mathbf{g}_B^n \geq 0 \quad (16)$$

$$\Rightarrow \boldsymbol{\Delta}_B^{*t}\mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* \leq \boldsymbol{\Delta}_B^{*t}\mathbf{g}_B^n \quad (17)$$

This gives us that

$$-\frac{1}{2}\boldsymbol{\Delta}_B^{*t}\mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* + \mathbf{g}_B^{nt}\boldsymbol{\Delta}_B^* \geq \frac{1}{2}\boldsymbol{\Delta}_B^{*t}\mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* \quad (18)$$

$$\Rightarrow f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) \geq \frac{1}{2}\boldsymbol{\Delta}_B^{*t}\mathbf{Q}_{BB}\boldsymbol{\Delta}_B^* \quad (19)$$

$$\Rightarrow f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) \geq \nu_B\frac{1}{2}\boldsymbol{\Delta}_B^{*t}\boldsymbol{\Delta}_B^* \quad (20)$$

where ν_B is the minimum eigenvalue of the matrix $\mathbf{Q}_{BB}$. Since $\mathbf{Q}_{BB}$ is positive definite always, this value is always greater than zero, and bounded below by the minimum eigenvalue among all 2×2 positive definite submatrices of $\mathbf{Q}$. Thus

$$f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) \geq \sigma\boldsymbol{\Delta}_B^{*t}\boldsymbol{\Delta}_B^* = \sigma\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n\|^2 \quad (21)$$

for some fixed $\sigma \geq 0$. □

Theorem 1. *The sequence $\{\boldsymbol{\theta}^n\}$ generated by Algorithm 1 converges.*

Proof. From Lemma 2, we have that $f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) \geq 0$. Thus the sequence $\{f(\boldsymbol{\theta}^n)\}$ is monotonically increasing. Since it is bounded from above (by the optimum value) it must converge. Since convergent sequences are Cauchy, this sequence is also Cauchy. Thus for every ϵ, $\exists n_0$ s.t $f(\boldsymbol{\theta}^{n+1}) - f(\boldsymbol{\theta}^n) \leq \sigma\epsilon^2\ \forall n \geq n_0$. Again using Lemma 2, we get that

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^n\|^2 \leq \epsilon^2 \quad (22)$$

for every $n \geq n_0$. Hence the sequence $\{\boldsymbol{\theta}^n\}$ is Cauchy. The feasible set of $\boldsymbol{\theta}$ is closed and compact, so Cauchy sequences are also convergent. Hence $\{\boldsymbol{\theta}^n\}$ converges. □

5 Finite termination

We have shown that $\{\boldsymbol{\theta}^n\}$ converges. Let $\hat{\boldsymbol{\theta}}$ be a limit point of $\{\boldsymbol{\theta}^n\}$. We will start from the assumption that the algorithm runs for an infinite number of iterations and then prove a contradiction.

Call the variable θ_{ik} as τ-violating if the magnitude of the projected gradient $\tilde{g}_{ik}$ is greater than τ. Note that at every iteration, the chosen set of variables contains at least one that is τ-violating. Now suppose the algorithm runs for an infinite number of iterations. Then it means that the sequence of iterates $\boldsymbol{\theta}^k$ contains an infinite number of τ-violating variables. Since there are only a finite number of distinct variables, we have that at least one variable figures as a τ-violating variable in the chosen set B an infinite number of times. Suppose that θ_{il} is one such variable, and let $\{k_{il}\}$ be the subsequence in which this variable is chosen as a τ-violating variable.

Lemma 3. *For every ϵ $\exists k_{il}^0$ s.t $\mid \theta_{il}^{k_{il}+1} - \theta_{il}^{k_{il}} \mid \leq \epsilon \quad \forall k_{il} > k_{il}^0$.*

Proof. We have that since $\boldsymbol{\theta}^k \rightarrow \hat{\boldsymbol{\theta}}$, $\boldsymbol{\theta}^{k_{il}} \rightarrow \hat{\boldsymbol{\theta}}$, and $\boldsymbol{\theta}^{k_{il}+1} \rightarrow \hat{\boldsymbol{\theta}}$. Thus, for any given ϵ $\exists$ k_{il}^0 such that

$$\mid \theta_{il}^{k_{il}} - \hat{\theta}_{il} \mid \leq \epsilon/2 \qquad \forall k_{il} > k_{il}^0 \tag{23}$$

$$\mid \theta_{il}^{k_{il}+1} - \hat{\theta}_{il} \mid \leq \epsilon/2 \qquad \forall k_{il} + 1 > k_{il}^0 \tag{24}$$

This gives, by triangle inequality,

$$\mid \theta_{il}^{k_{il}+1} - \theta_{il}^{k_{il}} \mid \leq \epsilon \qquad \forall k_{il} > k_{il}^0 \tag{25}$$

□

Lemma 4. *$\mid \hat{g_{il}} \mid \geq \tau$, where $\hat{g_{il}}$ is the derivative of f w.r.t θ_{il} at $\hat{\boldsymbol{\theta}}$.*

Proof. This is simply because of the fact that $\mid g_{il}^{k_{il}} \mid \geq \tau$ for every k_{il}, and the absolute value of the derivative w.r.t θ_{il} is a continuous function of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}^{k_{il}} \rightarrow \hat{\boldsymbol{\theta}}$. □

We use some notation. If $\theta_{il}^{k_{il}} \in (l_{il}, u_{il})$ and if $\theta_{il}^{k_{il}+1} = l_{il}$ or $\theta_{il}^{k_{il}+1} = u_{il}$, then we say that "$k_{il}$ is int $\rightarrow$ bd", where "int" stands for interior and "bd" stands for boundary. Similar interpretations are assumed for "bd $\rightarrow$ bd" and "int $\rightarrow$ int". Thus each iteration k_{il} can be of one of only four possible kinds: int $\rightarrow$ int,int $\rightarrow$ bd, bd $\rightarrow$ int and bd $\rightarrow$ bd. We will prove that each of these kinds of iterations can only occur a finite number of times.

Lemma 5. *There can be only a finite number of int $\rightarrow$ int and bd $\rightarrow$ int transitions.*

Proof. Suppose not. Then we can construct an infinite subsequence $\{s_{il}\}$ of the sequence $\{k_{il}\}$ that consists of these transitions. Then we have that $g_{il}^{s_{il}+1} = 0$, using Lemma 1. Hence $g_{il}^{s_{il}+1} \rightarrow 0$. Since the gradient is a continuous function of $\boldsymbol{\theta}$, and since $\boldsymbol{\theta}^{s_{il}+1} \rightarrow \hat{\boldsymbol{\theta}}$, we have that $g_{il}^{s_{il}+1} \rightarrow \hat{g}_{il}$. But this means $\hat{g}_{il} = 0$, which contradicts Lemma 4. □

Lemma 6. *There can be only a finite number of int $\rightarrow$ bd transitions.*

Proof. Suppose that we have completed sufficient number of iterations so that all int $\rightarrow$ int and bd $\rightarrow$ int transitions have completed. The next int $\rightarrow$ bd transition will place θ_{il} on the boundary. Since there are no bd $\rightarrow$ int transitions anymore, θ_{il} will stay on the boundary henceforth. Hence there can be no more int $\rightarrow$ bd transitions. □

Lemma 7. *There can only be a finite number of bd $\rightarrow$ bd transitions.*

Proof. Suppose not, i.e there are an infinite number of bd $\rightarrow$ bd transitions. Let t_{il} be the subsequence of k_{il} consisting of bd $\rightarrow$ bd transitions. Now, the sequence $\theta_{il}^{t_{il}} \rightarrow \hat{\theta}_{il}$ and is therefore Cauchy. Hence $\exists n_1$ s.t

$$\mid \theta_{il}^{t_{il}} - \theta_{il}^{t_{il}+1} \mid \leq \epsilon \ll u_{il} - l_{il} \qquad \forall t_{il} \geq n_1 \tag{26}$$

Similarly, because the gradient is a continuous function of θ, the sequence $\{g_{il}^{t_{il}}\}$ is convergent and therefore Cauchy. Hence $\exists n_2$ s.t

$$\mid g_{il}^{t_{il}} - g_{il}^{t_{il}+1} \mid \leq \frac{\tau}{2} \qquad \forall k_{il} \geq n_2 \tag{27}$$

Also, from the previous lemmas, $\exists n_3$ s.t t_{il} is not int $\rightarrow$ int, bd $\rightarrow$ int or int $\rightarrow$ bd $\forall t_{il} \geq n_3$.

Take $n_0 = \max(n_1, n_2, n_3)$. Now, consider $t_{il} \geq n_0$. Without loss of generality, assume that $\theta_{il}^{t_{il}} = l_{il}$. Then, since $\mid \tilde{g}_{il}^{t_{il}} \mid \geq \tau$, we must have that $g_{il}^{t_{il}} \geq \tau$. From (26), and using the fact that this is a bd $\rightarrow$ bd transition, we must have that

$$\theta_{il}^{t_{il}+1} = l_{il} \tag{28}$$

From (27), we have that

$$g_{il}^{t_{il}+1} \geq \frac{\tau}{2} \tag{29}$$

From (28) and (29), we have that $\tilde{g}_{il}^{t_{il}+1} \geq \frac{\tau}{2}$, which contradicts Lemma 1. □

But if all int $\rightarrow$ int, int $\rightarrow$ bd, bd $\rightarrow$ int and bd $\rightarrow$ bd transitions are finite, then θ_{il} cannot be τ-violating an infinite number of times and hence we have a contradiction. This gives us the following theorem:

Theorem 2. *Algorithm 1 terminates in finite number of steps*

A Optimality conditions

This section proves three results on quadratic programs that have been used in the proof above.

Consider the optimization problem:

$$\max_{\mathbf{x}} -\frac{1}{2}\mathbf{x}^t\mathbf{H}\mathbf{x} + \mathbf{p}^t\mathbf{x} \tag{30}$$

s.t

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$$

The feasible set of the above problem is defined by box constraints and is therefore convex. If the Hessian $\mathbf{H}$ is positive definite, then this optimization problem will be convex. The gradient of f is denoted as ∇f. The projected gradient of f, denoted as $\nabla^P f$ is given by:

$$\nabla_i^P f(\mathbf{x}) = \begin{cases} \nabla_i f(\mathbf{x}) & \text{if } l_i < x_i < u_i \\ \max(0, \nabla_i f(\mathbf{x})) & \text{if } x_i = l_i \\ \min(0, \nabla_i f(\mathbf{x})) & \text{if } x_i = u_i \end{cases} \tag{31}$$

Lemma 8. *Consider the optimization problem* (30)*. Suppose* $\mathbf{H}$ *is positive definite. Then,* $\mathbf{x}$ *is optimum for* (30) *iff*

1. $\mathbf{x}$ *is feasible*
2. $\nabla_i^P f(\mathbf{x}) = 0 \qquad \forall i$

Proof. First let's write it as a minimization problem:

$$\min_{\mathbf{x}} h(\mathbf{x}) = -f(\mathbf{x}) \tag{32}$$

s.t

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$$

The positive definiteness of $\mathbf{H}$ and the fact that the constraints are box constraints imply that the problem is convex. Because the problem is convex, and because strong duality holds, a point $\mathbf{x}$ is optimal iff the KKT conditions hold at $\mathbf{x}$. The Lagrangian of the above problem is:

$$\mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b}) = h(\mathbf{x}) + \mathbf{a}^t(\mathbf{l} - \mathbf{x}) + \mathbf{b}^t(\mathbf{x} - \mathbf{u}) \tag{33}$$

The KKT conditions are then:

$$\nabla_{\mathbf{x}}\mathcal{L} = \nabla h(\mathbf{x}) - \mathbf{a} + \mathbf{b} = \mathbf{0} \tag{34}$$

$$a_i(l_i - x_i) = 0 \tag{35}$$

$$b_i(u_i - x_i) = 0 \tag{36}$$

$$\mathbf{a} \geq 0 \tag{37}$$

$$\mathbf{b} \geq 0 \tag{38}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \tag{39}$$

From (34), (35) and (36), we get that:

$$\frac{\partial h}{\partial x_i}(\mathbf{x}) = a_i - b_i \qquad \forall i \tag{40}$$

$$\Rightarrow \frac{\partial h}{\partial x_i}(\mathbf{x}) = \begin{cases} 0 & \text{if } l_i < x_i < u_i \\ a_i \geq 0 & \text{if } l_i = x_i \\ -b_i \leq 0 & \text{if } x_i = u_i \end{cases} \qquad \forall i \tag{41}$$

$$\Rightarrow \frac{\partial f}{\partial x_i}(\mathbf{x}) = \begin{cases} 0 & \text{if } l_i < x_i < u_i \\ -a_i \leq 0 & \text{if } l_i = x_i \\ b_i \geq 0 & \text{if } x_i = u_i \end{cases} \quad \forall i \tag{42}$$

$$\Rightarrow \nabla_i^P f(\mathbf{x}) = 0 \qquad \forall i \tag{43}$$

Conversely, suppose $\nabla_i^P f(\mathbf{x}) = 0$ and $\mathbf{x}$ is feasible. Taking $a_i = \max(\frac{\partial h}{\partial x_i}(\mathbf{x}), 0)$ and $b_i = -\min(\frac{\partial h}{\partial x_i}(\mathbf{x}), 0)$, we have that

$$l_i < x_i < u_i \tag{44}$$

$$\Rightarrow \frac{\partial h}{\partial x_i}(\mathbf{x}) = 0 \tag{45}$$

$$\Rightarrow a_i = 0 \text{ and } b_i = 0 \tag{46}$$

$$x_i = l_i \tag{47}$$

$$\Rightarrow \frac{\partial h}{\partial x_i}(\mathbf{x}) = -\frac{\partial f}{\partial x_i}(\mathbf{x}) \geq 0 \tag{48}$$

$$\Rightarrow a_i \geq 0 \text{ and } b_i = 0 \tag{49}$$

$$x_i = u_i \tag{50}$$

$$\Rightarrow \frac{\partial h}{\partial x_i}(\mathbf{x}) = -\frac{\partial f}{\partial x_i}(\mathbf{x}) \leq 0 \tag{51}$$

$$\Rightarrow a_i = 0 \text{ and } b_i \geq 0 \tag{52}$$

Thus (34), (35), (36), (37) and (38) are satisfied by this choice of $\mathbf{a}$ and $\mathbf{b}$. Thus the KKT conditions are equivalent to the following conditions, which are therefore necessary and sufficient for $\mathbf{x}$ to be optimal:

$$\nabla_i^P f(\mathbf{x}) = 0 \qquad \forall i \tag{53}$$

$$\mathbf{l} \leq \mathbf{x} \leq \mathbf{u} \tag{54}$$

□

Lemma 9. *Consider the optimization problem (30). Suppose $\mathbf{H}$ is positive definite. If $\mathbf{x}^*$ is optimal, and $\mathbf{x} \neq \mathbf{x}^*$ is feasible, then $(\mathbf{x}^* - \mathbf{x})^t \nabla f(\mathbf{x}^*) \geq 0$*

Proof. Because the feasible set is convex and because both $\mathbf{x}$ and $\mathbf{x}^*$ are feasible, we have that $\mathbf{x}(\lambda) = \mathbf{x} + \lambda(\mathbf{x}^* - \mathbf{x})$ is feasible for all $\lambda \in [0, 1]$. Consider the optimization problem:

$$\max_\lambda f(\mathbf{x}(\lambda)) - f(\mathbf{x}) \tag{55}$$

s.t

$$0 \leq \lambda \leq 1$$

From a Taylor series expansion, it can be seen that

$$f(\mathbf{x}(\lambda)) - f(\mathbf{x}) = -\frac{\lambda^2}{2}(\mathbf{x}^* - \mathbf{x})^t H (\mathbf{x}^* - \mathbf{x}) + \lambda(\mathbf{x}^* - \mathbf{x})^t \nabla f(\mathbf{x}) \tag{56}$$

Since $\mathbf{H}$ is positive definite,$(\mathbf{x}^* - \mathbf{x})^t \mathbf{H} (\mathbf{x}^* - \mathbf{x}) > 0$. Hence, the problem (55) satisfies the conditions of Lemma 8. Also, because $\mathbf{x}^*$ is optimal for (30), $\lambda = 1$ is optimal for (55). Hence, by Lemma 8, the projected gradient of (55) at $\lambda = 1$ is 0. This means that

$$\left. \frac{\partial f(\mathbf{x}(\lambda))}{\partial \lambda} \right|_{\lambda=1} \geq 0 \tag{57}$$

This gives, using the chain rule

$$\left. (\mathbf{x} - \mathbf{x}^*)^t \nabla f(\mathbf{x}(\lambda)) \right|_{\lambda=1} \geq 0 \tag{58}$$

$$\Rightarrow (\mathbf{x}^* - \mathbf{x})^t \nabla f(\mathbf{x}^*) \geq 0 \tag{59}$$

where the last step uses the fact that $\mathbf{x}(1) = \mathbf{x}^*$. □

Lemma 10. *Consider the optimization problem (30). Suppose* $\mathbf{H}$ *is positive definite. If* $\mathbf{x}$ *is feasible and* $|\nabla^P_{ik} f(\mathbf{x})| \leq \tau$*, and if* f^* *is the optimal value, then*

$$f^* - f(\mathbf{x}) \leq \frac{d^2\tau^2}{2m} \tag{60}$$

where d *is the dimensionality of* $\mathbf{x}$ *and* m *is the least eigenvalue of* $\mathbf{H}$*.*

Proof. This proof is the extension of the proof in [1, p. 459]. Again, take $h = -f$ and define the Lagrangian. Then, for the given $\mathbf{x}$, we can define

$$\hat{a}_i = \max(\frac{\partial h}{\partial x_i}(\mathbf{x}), 0) \tag{61}$$

$$a_i = \begin{cases} \hat{a}_i & \text{if } \hat{a}_i \geq \tau \\ 0 & ow \end{cases} \tag{62}$$

$$\hat{b}_i = -\min(\frac{\partial h}{\partial x_i}(\mathbf{x}), 0) \tag{63}$$

$$b_i = \begin{cases} \hat{b}_i & \text{if } \hat{b}_i \geq \tau \\ 0 & ow \end{cases} \tag{64}$$

Then, equations (35), (36), (37) and (38) are satisfied by this choice of $\mathbf{a}$ and $\mathbf{b}$. Equations (35) and (36) mean that this choice of $\mathbf{a}$ and $\mathbf{b}$ satisfies:

$$h(\mathbf{x}) = \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b}) \tag{65}$$

(34) is only satisfied to a precision of τ, which means that

$$\|\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b})\| \leq d\tau \tag{66}$$

Also note that:

$$\begin{aligned} h^* = -f^* &= \max_{\mathbf{a},\mathbf{b}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b}) \\ &\geq \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b}) \\ &= \mathcal{L}(\mathbf{x}^0, \mathbf{a}, \mathbf{b}) \end{aligned} \tag{67}$$

for some $\mathbf{x}^0$.

Hence,

$$h(\mathbf{x}) - h^* \leq h(\mathbf{x}) - \mathcal{L}(\mathbf{x}^0, \mathbf{a}, \mathbf{b}) \tag{68}$$

$$\Rightarrow h(\mathbf{x}) - h^* \leq \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b}) - \mathcal{L}(\mathbf{x}^0, \mathbf{a}, \mathbf{b}) \tag{69}$$

Using a Taylor series expansion of $g(\mathbf{x}) = \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b})$ we see that:

$$\begin{aligned} g(\mathbf{x}^0) - g(\mathbf{x}) &= (\mathbf{x}^0 - \mathbf{x})^t \nabla g(\mathbf{x}) \\ &+ \frac{1}{2}(\mathbf{x}^0 - \mathbf{x})^t \nabla^2 g(\mathbf{x})(\mathbf{x}^0 - \mathbf{x}) \end{aligned} \tag{70}$$

The Hessian $\nabla^2 g(\mathbf{x})$ is nothing but $\mathbf{H}$. The quantity $(\mathbf{x}^0-\mathbf{x})^t \nabla^2 g(\mathbf{x})(\mathbf{x}^0-\mathbf{x})$ is lowerbounded by $m\|\mathbf{x}^0 - \mathbf{x}\|^2$, $m > 0$ being the least eigenvalue of $\mathbf{H}$. Hence,

$$g(\mathbf{x}^0) - g(\mathbf{x}) \geq (\mathbf{x}^0 - \mathbf{x})^t \nabla g(\mathbf{x}) + \frac{m}{2}\|\mathbf{x}^0 - \mathbf{x}\|^2 \tag{71}$$

The right hand side is a convex function of $\mathbf{x}^0$ for fixed $\mathbf{x}$. Setting the gradient $= 0$, we find that the right hand side attains its minimum value of $-\frac{1}{2m}\|\nabla g(\mathbf{x})\|^2$ at $\mathbf{x}^0 = \mathbf{x} - \frac{1}{m}\nabla g(\mathbf{x})$. Thus,

$$g(\mathbf{x}^0) - g(\mathbf{x}) \geq -\frac{1}{2m}\|\nabla g(\mathbf{x})\|^2 \tag{72}$$

Combining (66), (69) and (72), and noting that $f = -h$ and $g(\mathbf{x}) = \mathcal{L}(\mathbf{x}, \mathbf{a}, \mathbf{b})$ we get the desired inequality. □

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[2] B. Hariharan, M. Varma, S. V. N. Vishwanathan, and L. Zelnik-Manor. Max-margin multi-label classification with correlations. In *Proceedings of the International Conference on Machine Learning (To appear)*, 2010.

[3] S. S. Keerthi and E. G. Gilbert. Convergence of a generalized smo algorithm for svm classifier design. In *Machine Learning*, 2000.