Target Neighbor Consistent Feature Weighting for Nearest Neighbor Classification (Supplementary Material)

Ichiro Takeuchi
Department of Engineering
Nagoya Institute of Technology
takeuchi.ichiro@nitech.ac.jp

Masashi Sugiyama
Department of Computer Science
Tokyo Institute of Technology
sugi@cs.titech.ac.jp

In supplement A, we describe how to solve the equality constrained QP (4). In supplements B, C and D, the proofs of Lemmas 5, 6 and Theorem 7 are presented, respectively. We explain the finite termination property of the SQP-FW algorithm in supplement E, while regularization path tracking algorithm is described in supplement F. Finally, an illustrative example of the algorithm behavior is presented in supplement G.

In this supplement, we use the following notations as well as those used in the main text. To denote sub-vectors or sub-matrices, we use index sets. For example, $v_{\mathcal{A}}$ for a vector v and index set $\mathcal{A}$ indicates a sub-vector of v having only the elements in an index set $\mathcal{A}$. Similarly, $M_{\mathcal{A},\mathcal{B}}$ indicates a sub-matrix of M having only the rows in $\mathcal{A}$ and the columns in $\mathcal{B}$. We also use notation such as $M_{\mathcal{A},:}$ to denote a sub-matrix of M having only the rows in $\mathcal{A}$. In addition, the $n \times n$ identity matrix is denoted by I_n.

A How to solve the equality constrained QP (4)

We now describe how to solve the equality constrained QP (EQP) (4). We first note that, for each instance $i \in \mathbb{N}_n$, there exists at least one $(i,h) \in \mathcal{H}^{[=]}$ and $(i,m) \in \mathcal{M}^{[=]}$. We divide $\mathcal{H}^{[=]}$ into $\tilde{\mathcal{H}}^{[=]}$ and $\hat{\mathcal{H}}^{[=]}$ so that $\tilde{\mathcal{H}}^{[=]}$ has exactly one (i,h) for each $i \in \mathbb{N}_n$, and $\hat{\mathcal{H}}^{[=]}$ has the rest of $(i,h) \in \mathcal{H}^{[=]}$. Note that $|\tilde{\mathcal{H}}^{[=]}| = n$ and $|\hat{\mathcal{H}}^{[=]}| = |\mathcal{H}^{[=]}| - n$. Similarly, $\mathcal{M}^{[=]}$ is divided into $\tilde{\mathcal{M}}^{[=]}$ and $\hat{\mathcal{M}}^{[=]}$. The EQP (4) is written as follows:

$$\min_{\Delta w, \Delta \xi, \Delta \eta} \quad \theta n^{-1} \mathbf{1}^{\top}((\xi + \Delta \xi) - (\eta + \Delta \eta)) + \frac{1}{2}\|(w + \Delta w) - \bar{w}\|_2^2 \tag{6a}$$

$$\text{s.t.} \quad \mathbf{1}^{\top}(w + \Delta w) = 1, \tag{6b}$$

$$w_{\mathcal{Z}} + \Delta w_{\mathcal{Z}} = \mathbf{0}, \tag{6c}$$

$$\tilde{D}(w + \Delta w) - I_n(\xi + \Delta \xi) = \mathbf{0}, \tag{6d}$$

$$\hat{D}(w + \Delta w) - \hat{E}(\xi + \Delta \xi) = \mathbf{0}, \tag{6e}$$

$$\tilde{F}(w + \Delta w) - I_n(\eta + \Delta \eta) = \mathbf{0}, \tag{6f}$$

$$\hat{F}(w + \Delta w) - \hat{G}(\eta + \Delta \eta) = \mathbf{0}, \tag{6g}$$

where $\tilde{D}$ is the $n \times \ell$ matrix whose rows are $\varepsilon_{i,h}^{\top}$ for $(i,h) \in \tilde{H}^{[=]}$, $\hat{D}$ is the $(|\mathcal{H}^{[=]}| - n) \times \ell$ matrix whose rows are $\varepsilon_{i,h}^{\top}$ for $(i,h) \in \hat{H}^{[=]}$, and $\hat{E}$ is the $(|\mathcal{H}^{[=]}| - n) \times n$ matrix whose rows are unit vectors having 1 at the corresponding i-th element and 0s at the other elements. In addition to computing the solutions $(\Delta w, \Delta \xi, \Delta \eta)$, we need to compute the Lagrange multipliers for each of the equality constraints. The Lagrange multipliers corresponding to (6b), (6c), (6d), (6e), (6f), and

(6g), are denoted by $\phi \in \mathbb{R}$, $\alpha \in \mathbb{R}^{|\mathcal{Z}|}$, $\tilde{\beta} \in \mathbb{R}^{|\tilde{\mathcal{H}}^{[=]}|}$, $\hat{\beta} \in \mathbb{R}^{|\hat{\mathcal{H}}^{[=]}|}$, $\tilde{\gamma} \in \mathbb{R}^{|\tilde{\mathcal{M}}^{[=]}|}$ and $\hat{\gamma} \in \mathbb{R}^{|\hat{\mathcal{M}}^{[=]}|}$, respectively[7].

The EQP (6) is greatly simplified by eliminating some variables. First, from (6d) and (6f),

$$\Delta \xi \;=\; \tilde{D}\Delta w + (\tilde{D}w - \xi) = \tilde{D}\Delta w, \tag{7a}$$

$$\Delta \eta \;=\; \tilde{F}\Delta w + (\tilde{D}w - \eta) = \tilde{F}\Delta w. \tag{7b}$$

Furthermore, from (6c),

$$\Delta w_{\mathcal{Z}} = -w_{\mathcal{Z}} = \mathbf{0}. \tag{8}$$

By eliminating these variables and substituting

$$
\begin{aligned}
\mathbf{1}^\top w &= 1, \\
w_{\mathcal{Z}} &= \mathbf{0}, \\
\hat{D}w - \hat{E}\xi &= \mathbf{0}, \\
\hat{F}w - \hat{G}\eta &= \mathbf{0},
\end{aligned}
$$

the EQP (6) is simplified as

$$\min_{\Delta w_{\mathcal{P}}} \quad \theta n^{-1}\mathbf{1}^\top((\xi - \eta) + (\tilde{D}_{:,\mathcal{P}} - \tilde{F}_{:,\mathcal{P}})\Delta w_{\mathcal{P}})) + \frac{1}{2}(\|w_{\mathcal{P}} + \Delta w_{\mathcal{P}} - \bar{w}_{\mathcal{P}}\|_2^2 + \|\bar{w}_{\mathcal{Z}}\|_2^2) \tag{9a}$$

$$\text{s.t.} \quad \mathbf{1}^\top \Delta w_{\mathcal{P}} = 0, \tag{9b}$$

$$(\hat{D} - \hat{E}\tilde{D}_{:,\mathcal{P}})\Delta w_{\mathcal{P}} = \mathbf{0}, \tag{9c}$$

$$(\hat{F} - \hat{G}\tilde{F}_{:,\mathcal{P}})\Delta w_{\mathcal{P}} = \mathbf{0}. \tag{9d}$$

The KKT optimality conditions of (9) are written as

$$\Delta w_{\mathcal{P}} \;+\; \left[\begin{array}{ccc} \mathbf{1} & ((\hat{D} - \hat{E}\tilde{D})_{:,\mathcal{P}})^\top & ((\hat{F} - \hat{G}\tilde{F})_{:,\mathcal{P}})^\top \end{array}\right] \left[\begin{array}{c} \phi \\ \hat{\beta} \\ \hat{\gamma} \end{array}\right]$$

$$= \; -(w_{\mathcal{P}} - \bar{w}_{\mathcal{P}}) - (\tilde{D}_{:,\mathcal{P}} - \tilde{F}_{:,\mathcal{P}})^\top \mathbf{1}\theta, \tag{10a}$$

$$\left[\begin{array}{c} \mathbf{1}^\top \\ (\hat{D} - \hat{E}\tilde{D})_{:,\mathcal{P}} \\ (\hat{F} - \hat{G}\tilde{F})_{:,\mathcal{P}} \end{array}\right] \Delta w_{\mathcal{P}} \;=\; \mathbf{0}. \tag{10b}$$

Multiplying (10a) by the matrix in (10b) from the left and subtracting (10b) yields the following system of linear equations:

$$\left[\begin{array}{ccc} |\mathcal{P}| & \mathbf{1}^\top((\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}})^\top & \mathbf{1}^\top((\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}})^\top \\ (\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}}\mathbf{1} & (\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}}((\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}})^\top & (\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}}((\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}})^\top \\ (\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}}\mathbf{1} & (\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}}((\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}})^\top & (\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}}((\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}})^\top \end{array}\right] \left[\begin{array}{c} \phi \\ \hat{\beta} \\ \hat{\gamma} \end{array}\right]$$

$$= \left[\begin{array}{c} -\mathbf{1}^\top(w_{\mathcal{P}} - \bar{w}_{\mathcal{P}}) \\ -(\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}}(w_{\mathcal{P}} - \bar{w}_{\mathcal{P}}) \\ -(\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}}(w_{\mathcal{P}} - \bar{w}_{\mathcal{P}}) \end{array}\right] + \left[\begin{array}{c} -\mathbf{1}^\top((\tilde{D}-\tilde{F})_{:,\mathcal{P}})^\top\mathbf{1} \\ -(\hat{D}-\hat{E}\tilde{D})_{:,\mathcal{P}}((\tilde{D}-\tilde{F})_{:,\mathcal{P}})^\top\mathbf{1} \\ -(\hat{F}-\hat{G}\tilde{F})_{:,\mathcal{P}}((\tilde{D}-\tilde{F})_{:,\mathcal{P}})^\top\mathbf{1} \end{array}\right] \theta. \tag{11}$$

After obtaining $(\phi, \hat{\beta}, \hat{\gamma})$ by solving (11), $\Delta w_{\mathcal{P}}$ is computed as

$$\Delta w_{\mathcal{P}} = \; -\left[\begin{array}{ccc} \mathbf{1} & ((\hat{D} - \hat{E}\tilde{D})_{:,\mathcal{P}})^\top & ((\hat{F} - \hat{G}\tilde{F})_{:,\mathcal{P}})^\top \end{array}\right] \left[\begin{array}{c} \phi \\ \hat{\beta} \\ \hat{\gamma} \end{array}\right]$$

$$-(w_{\mathcal{P}} - \bar{w}_{\mathcal{P}}) - ((\tilde{D} - \tilde{F})_{:,\mathcal{P}})^\top \mathbf{1}\theta, \tag{12}$$

---

[7]Here, we need to clearly mention how the Lagrange multipliers are defined because the signs of these multipliers play important roles in our algorithm (see Lemma 6). They are defined in such a way that the Lagrangian of the problem (4) is written as $L := (\text{loss term}) + \phi(\mathbf{1}^\top(w + \Delta w) - 1) + \sum_{j \in \mathcal{Z}} \alpha_j(w_j + \Delta w_j) + \sum_{(i,h) \in \mathcal{H}^{[=]}} \beta_{(i,h)}(\varepsilon_{i,h}^\top(w + \Delta w) - (\xi_i + \Delta \xi_i)) + \sum_{(i,m) \in \mathcal{M}^{[=]}} \gamma_{(i,m)}(\varepsilon_{i,m}^\top(w + \Delta w) - (\eta_i + \Delta \eta_i))$.

and $(\Delta\xi, \Delta\eta)$ are computed from (7). The rest of the Lagrange multipliers $(\alpha, \tilde{\beta}, \tilde{\gamma})$ are computed as

$$\alpha = \bar{w}_{\mathcal{Z}} - \mathbf{1}\phi - (\tilde{D}_{:,\mathcal{Z}})^\top \tilde{\beta} - (\hat{D}_{:,\mathcal{Z}})^\top \hat{\beta} - (\tilde{F}_{:,\mathcal{Z}})^\top \tilde{\gamma} - (\hat{F}_{:,\mathcal{Z}})^\top \hat{\gamma}, \tag{13a}$$

$$\tilde{\beta} = \mathbf{1}\theta - \hat{E}^\top \hat{\beta}, \tag{13b}$$

$$\tilde{\gamma} = \mathbf{1}\theta - \hat{G}^\top \hat{\gamma}. \tag{13c}$$

A.1 Computational complexity

In the above computation, we divided $\mathcal{H}^{[=]}$ into $\tilde{\mathcal{H}}^{[=]}$ and $\hat{\mathcal{H}}^{[=]}$, where $|\tilde{\mathcal{H}}^{[=]}| = n$ and $|\hat{\mathcal{H}}^{[=]}| = |\mathcal{H}^{[=]}| - n$ (similarly for $\mathcal{M}^{[=]}$, $\tilde{\mathcal{M}}^{[=]}$ and $\hat{\mathcal{M}}^{[=]}$). Since the linear system of equations (11) has $1 + |\hat{\mathcal{H}}^{[=]}| + |\hat{\mathcal{M}}^{[=]}|$ unknowns and equations, we briefly discuss the sizes of $\hat{\mathcal{H}}^{[=]}$ and $\hat{\mathcal{M}}^{[=]}$. First, note that, if we have exactly one target hit for each instance $i \in \mathbb{N}_n$, the size of $\mathcal{H}^{[=]}$ is n and thus $\hat{\mathcal{H}}^{[=]}$ is empty. In other words, $\hat{\mathcal{H}}^{[=]}$ contains index pairs (i, h) such that the instance i has more than one target hits (one of them is in $\tilde{\mathcal{H}}^{[=]}$ and the rest are in $\hat{\mathcal{H}}^{[=]}$). If the instance i has two target hits, say, h_1 and h_2, it means that $d(x_i, x_{h_1}|w) = d(x_i, x_{h_2}|w)$ and they both are tied κ nearest instance in $\mathbb{H}_i$. Although such tied target hits/misses frequently happens in the algorithm (this is one of the important mechanisms for handling TN change in the algorithm), the number of those tied TNs is usually very small at each step. Our experience indicates that it is irrespective to the sample size n and $\mathcal{O}(1)$. Therefore, the main computational cost for computing $(\Delta w, \Delta\xi, \Delta\eta)$ is in matrix-vector multiplications involving $n \times |\mathcal{P}|$ matrices, while the cost for computing (α, β, γ) is in matrix-vector multiplications involving $n \times |\mathcal{Z}|$ matrices.

A.2 Strict convexity and linear independence in the active set constraints

The objective function of the reduced EQP (9) is strictly convex w.r.t. the variables $\Delta w_{\mathcal{P}}$ because the quadratic term of the objective function is $(\Delta w_{\mathcal{P}})^\top I_{|w_{\mathcal{P}}|}(\Delta w_{\mathcal{P}})$. In addition, we can show that the equality constraints (9b), (9c) and (9d) are always linearly independent. Taken together, the problem (9) always has a unique solution and the matrix in (11) is always nonsingular and its Cholesky decomposition exists. The linear dependency of the equality constraints is confirmed by the following argument. First, we can always start the algorithm with empty $\hat{\mathcal{H}}^{[=]}$ and $\hat{\mathcal{M}}^{[=]}$, in which there is only one equality constraint (9b) and it is thus clearly independent. (Even if there are tied TNs by coincidence in the initial weighted feature space by $\bar{w}$, we can just pick up one of them as the target hit/miss. The tied ones will be included later to $\hat{\mathcal{H}}^{[=]}$ or $\hat{\mathcal{M}}^{[=]}$ if needed.) Our algorithm adds up equality constraint one by one into the active set using Lemma 5. It can be shown that the constraint normal of the blocking constraint in Lemma 5 cannot be linear combination of the normals in the current active constraints (see page 476 in [15]). Therefore, linear independence is maintained after the blocking constraint is added to the working set. On the other hand, deletion of an index from the active set cannot introduce linear dependence.

B Proof of Lemma 5

Noting that $(\Delta x, \Delta\xi, \Delta\eta)$ is feasible (and TN-weight consistent) descent direction of the original problem (2) and $\tau = 1$ solves the EQP (4) over the current active set, the step length τ must be in $[0, 1]$. Since the step length τ should be determined so that it satisfies all the currently inactive constraints:

$$\begin{aligned}
w_j + \tau\Delta w_j &\geq 0, \text{ for } j \in \mathcal{P}, \\
\varepsilon_{i,h}^\top(w + \tau\Delta w) - (\xi_i + \tau\Delta\xi_i) &\leq 0, \text{ for } (i, h) \in \mathcal{H}^{[<]}, \\
\varepsilon_{i,h}^\top(w + \tau\Delta w) - (\xi_i + \tau\Delta\xi_i) &\geq 0, \text{ for } (i, h) \in \mathcal{H}^{[>]}, \\
\varepsilon_{i,m}^\top(w + \tau\Delta w) - (\eta_i + \tau\Delta\eta_i) &\leq 0, \text{ for } (i, m) \in \mathcal{M}^{[<]}, \\
\varepsilon_{i,m}^\top(w + \tau\Delta w) - (\eta_i + \tau\Delta\eta_i) &\geq 0, \text{ for } (i, m) \in \mathcal{M}^{[>]},
\end{aligned}$$

the maximum step length τ is determined as in (5).

C   Proof of Lemma 6

The conditions on the sizes of index sets $\{(\mathcal{H}_i^{[<]}, \mathcal{H}_i^{[=]}, \mathcal{H}_i^{[>]}, \mathcal{M}_i^{[<]}, \mathcal{M}_i^{[=]}, \mathcal{M}_i^{[>]})\}_{i \in \mathbb{N}_n}$ are required for TN-weight consistency. In order to satisfy TN-weight consistency, we need to guarantee that the κ nearest instance in $\mathbb{H}_i$ must be the member of $\mathcal{H}_i^{[=]}$. This condition is satisfied only when $|\mathcal{H}_i^{[<]}| \leq \kappa - 1$, $|\mathcal{H}_i^{[=]}| \geq 1$, and $|\mathcal{H}_i^{[>]}| \leq |\mathbb{H}_i| - \kappa$. Therefore, an element in $\mathcal{H}_i^{[=]}$ can move to $\mathcal{H}_i^{[<]}$ only when $|\mathcal{H}_i^{[<]}| \leq \kappa - 2$ and $|\mathcal{H}_i^{[=]}| \geq 2$, and an element in $\mathcal{H}_i^{[=]}$ can move to $\mathcal{H}_i^{[>]}$ only when $|\mathcal{H}_i^{[>]}| < |\mathbb{H}_i| - \kappa$ and $|\mathcal{H}_i^{[=]}| \geq 2$. Similarly, the λ nearest instance in $\mathbb{M}_i$ must be the member of $\mathcal{M}_i^{[=]}$. This condition is satisfied only when $|\mathcal{M}_i^{[<]}| \leq \lambda - 1$, $|\mathcal{M}_i^{[=]}| \geq 1$, and $|\mathcal{M}_i^{[>]}| \leq |\mathbb{M}_i| - \lambda$. Therefore, an element in $\mathcal{M}_i^{[=]}$ can move to $\mathcal{M}_i^{[<]}$ only when $|\mathcal{M}_i^{[<]}| \leq \lambda - 2$ and $|\mathcal{M}_i^{[=]}| \geq 2$, and an element in $\mathcal{M}_i^{[=]}$ can move to $\mathcal{M}_i^{[>]}$ only when $|\mathcal{M}_i^{[>]}| < |\mathbb{M}_i| - \lambda$ and $|\mathcal{M}_i^{[=]}| \geq 2$. Note that the algorithm allows $\mathcal{H}_i^{[=]}$ and $\mathcal{M}_i^{[=]}$ to have more than one tied target hits and misses, respectively.

On the other hand, the conditions on the signs of the Lagrange multipliers (α, β, γ) are required for guaranteeing strict decrease of the objective function. Suppose, for example, that our current solution (w, ξ, η) satisfies $d(x_i, x_h|w) - \xi_i = 0$ for a certain $(i, h) \in \mathcal{H}^{[=]}$. Lagrange multiplier theory tells that the objective function can be decreased by releasing the equality constraint toward $d(x_i, x_h|w) - \xi_i < 0$ if and only if the corresponding Lagrange multiplier $\beta_{(i,h)}$ is strictly negative. Similarly, the objective function can be decreased by releasing the equality constraint toward $d(x_i, x_h|w) - \xi_i > 0$ if and only if the corresponding Lagrange multiplier $\beta_{(i,h)}$ is strictly positive. See, for example, section 12.3 and Theorem 16.5 in [15] for Lagrange multiplier theory.

Taken together, the objective function in (2) can be decreased in the neighborhood while satisfying the feasibility and the TN-weight consistency if and only if the conditions of one of the rules in Lemma 6 are satisfied and the index sets $(\mathcal{H}, \mathcal{M}, \mathcal{Z}, \mathcal{P})$ are updated by the corresponding rule.

D   Proof of Theorem 7

Theorem 7 is direct consequence of Lemma 6. Optimality condition of convex QP indicates that, if $(\Delta w, \Delta \xi, \Delta \eta) = \mathbf{0}$, the current solution is the optimal solution over the current active set (which also means that it is optimal over current TNs). If none of the rules in Lemma 6 are applied to the current $(\mathcal{H}, \mathcal{M}, \mathcal{Z}, \mathcal{P})$ and (α, β, γ), there are no feasible and TN-weight consistent descent direction in the neighborhood of the current solution. It means that the current solution is the local optimal solution of (2).

E   Finite Termination Property

It can be shown that the SQP-FW algorithm converges to a local minimum solution in a finite number of iterations. Here, we explain this property following pages 477–478 in [15]. When $(\Delta w, \Delta \xi, \Delta \eta) \neq \mathbf{0}$, $(\Delta w, \Delta \xi, \Delta \eta)$ is shown to be a strict descent direction, and the objective function always strictly decreases if a nonzero step (i.e., $\tau > 0$) is taken (see Theorem 16.6 in [15]). On the other hand, when $(\Delta w, \Delta \xi, \Delta \eta) = \mathbf{0}$, the current solution is already the optimal solution for the current active set. If any of the rules in Lemma 6 are applied, the next step after dropping the corresponding active constraint is shown to be a non-zero step that strictly decreases the objective function (see Theorem 16.5 in [15]). Finally, when the zero step is taken ($\tau = 0.0$) under $(\Delta w, \Delta \xi, \Delta \eta) \neq \mathbf{0}$, the zero steps continue at most N times (where N is the variable dimension) because a new linearly independent constraint is added to the active set in each of such steps and thus the active set can have at most N linearly independent constraints. If the active set has the maximum N constraints, the solution to the EQP must be $(\Delta w, \Delta \xi, \Delta \eta) = \mathbf{0}$. These arguments indicate that the algorithm finds the optimal solution over the current active set at least once every N iterations. Since the number of possible active sets is finite and the algorithm never encounters the same active set, it terminates with a solution in a finite number of iterations.

F Regularization Path Tracking Algorithm

We describe the regularization path tracking algorithm for computing a path of solutions that satisfy the optimality condition in Theorem 7 for a range of regularization parameter θ.

Suppose we have a local optimal solution w that satisfies the optimality condition in Theorem 7 as well as the set of consistent TNs $\{(h_i^\kappa, m_i^\lambda)\}_{i \in \mathbb{N}_n}$ and the index sets $\{\mathcal{H}, \mathcal{M}, \mathcal{P}, \mathcal{Z}\}$ when $\theta = \theta_t$. In order to develop the regularization path tracking algorithm, we need to investigate how the optimality condition changes when we perturb the regularization parameter θ in the neighborhood of θ_t. Remembering that the problem (2) is formulated as a convex QP as long as the TNs are unchanged, we need to study the Karush-Khun-Tucker (KKT) optimality condition of the current solution at $\theta = \theta_t$ as in supplement A.

From (11), we can formulate $(\phi, \hat{\beta}, \hat{\gamma})$ as affine functions of θ. Since the solutions $(\Delta w, \Delta \xi, \Delta \eta)$ and the rest of the Lagrange multipliers $(\alpha, \tilde{\beta}, \tilde{\gamma})$ are affine functions of these $(\phi, \hat{\beta}, \hat{\gamma})$ and θ, they are all represented as affine functions of θ. Let us denote those functional relationships by

$$\Delta w = \mu^{(w)} + \nu^{(w)}\theta, \tag{14a}$$
$$\Delta \xi = \mu^{(\xi)} + \nu^{(\xi)}\theta, \tag{14b}$$
$$\Delta \eta = \mu^{(\eta)} + \nu^{(\eta)}\theta, \tag{14c}$$
$$\alpha = \mu^{(\alpha)} + \nu^{(\alpha)}\theta, \tag{14d}$$
$$\beta = \mu^{(\beta)} + \nu^{(\beta)}\theta, \tag{14e}$$
$$\gamma = \mu^{(\gamma)} + \nu^{(\gamma)}\theta, \tag{14f}$$

where $\mu^{(w,)} \ldots, \mu^{(\gamma)}$ and $\nu^{(w)} \ldots, \nu^{(\gamma)}$ are vectors with appropriate sizes computed from (11), (12) and (13).

Equations (14) imply that the optimal weights w changes linearly unless the active set (including those for TN-weight consistency) is changed. When we increase the regularization parameter θ from θ_t, we will encounter a point at which the active set must be updated. Those points are called *breakpoints* in the literature. At breakpoints, the linearity of the solutions breaks, where we update the active set and recompute the affine functions in (14) using the new active set. In order to detect the next breakpoint, we need to know when and which constraint enters to or exits from the active set.

In order to see when each inactive constraint moves into active set, we can use the fact that these inactive constraints can also be represented as affine functions of θ as follows:

$$w_\mathcal{P} + \Delta w_\mathcal{P} \geq 0$$
$$\iff (w_\mathcal{P} + \mu_\mathcal{P}^{(w)}) + \nu_\mathcal{P}^{(w)}\theta \geq 0, \tag{15a}$$
$$d(x_i, x_h | w + \Delta w) \leq \xi_i + \Delta \xi_i$$
$$\iff (\varepsilon_{i,h}^\top w - \xi_i + \varepsilon_{i,h}^\top \mu^{(w)} - \mu_i^{(\xi)}) + (\varepsilon_{i,h}^\top \nu^{(w)} - \nu_i^{(\xi)})\theta \leq 0 \text{ for } (i,h) \in \mathcal{H}^{[<]} \tag{15b}$$
$$d(x_i, x_h | w + \Delta w) \geq \xi_i + \Delta \xi_i$$
$$\iff (\varepsilon_{i,h}^\top w - \xi_i + \varepsilon_{i,h}^\top \mu^{(w)} - \mu_i^{(\xi)}) + (\varepsilon_{i,h}^\top \nu^{(w)} - \nu_i^{(\xi)})\theta \geq 0 \text{ for } (i,h) \in \mathcal{H}^{[>]} \tag{15c}$$
$$d(x_i, x_m | w + \Delta w) \leq \eta_i + \Delta \eta_i$$
$$\iff (\varepsilon_{i,m}^\top w - \eta_i + \varepsilon_{i,m}^\top \mu^{(w)} - \mu_i^{(\eta)}) + (\varepsilon_{i,m}^\top \nu^{(w)} - \nu_i^{(\eta)})\theta \leq 0 \text{ for } (i,m) \in \mathcal{M}^{[<]} \tag{15d}$$
$$d(x_i, x_m | w + \Delta w) \geq \eta_i + \Delta \eta_i$$
$$\iff (\varepsilon_{i,m}^\top w - \eta_i + \varepsilon_{i,m}^\top \mu^{(w)} - \mu_i^{(\eta)}) + (\varepsilon_{i,m}^\top \nu^{(w)} - \nu_i^{(\eta)})\theta \geq 0 \text{ for } (i,m) \in \mathcal{M}^{[>]}. \tag{15e}$$

Using these affine functions, we can identify the largest possible update of θ at which one of the inactive constraints becomes active.

On the other hand, we can also identify the active constraint that first goes out of the active set by examining the Lagrange multipliers in (14d), (14e) and (14f) as well as the sizes of $\{\mathcal{H}_i^{[<]}, \mathcal{H}_i^{[=]}, \mathcal{H}_i^{[>]}, \mathcal{M}_i^{[<]}, \mathcal{M}_i^{[=]}, \mathcal{M}_i^{[>]}\}_{i \in \mathbb{N}_n}$. Here, note that, TN-weight consistency requires that we can move out active constraints in $\mathcal{H}^{[=]}$ and $\mathcal{M}^{[=]}$ only when the κ-th nearest instance in $\mathbb{H}_i$ can stay in $\mathcal{H}_i^{[=]}$ and the λ-th nearest instance in $\mathbb{M}_i$ can stay in $\mathcal{M}_i^{[=]}$ after the movement.

Taken together, the next breakpoint θ_{t+1} is determined as

$$\theta_{t+1} = \min_{\theta \geq \theta_t} \Big\{ \quad \min_{j \in \mathcal{P}, \nu_j^{(w)} < 0} -\frac{w_j + \mu_j^{(w)}}{\nu_j^{(w)}}, \tag{16a}$$

$$\min_{(i,h) \in \mathcal{H}^{[<]}, \, \varepsilon_{i,h}^\top \nu^{(w)} > \nu_i^{(\xi)}} -\frac{\varepsilon_{i,h}^\top w - \xi_i + \varepsilon_{i,h}^\top \mu^{(w)} - \mu_j^{(\xi)}}{\varepsilon_{i,h}^\top \nu^{(w)} - \nu_i^{(\xi)}}, \tag{16b}$$

$$\min_{(i,h) \in \mathcal{H}^{[>]}, \, \varepsilon_{i,h}^\top \nu^{(w)} < \nu_i^{(\xi)}} -\frac{\varepsilon_{i,h}^\top w - \xi_i + \varepsilon_{i,h}^\top \mu^{(w)} - \mu_j^{(\xi)}}{\varepsilon_{i,h}^\top \nu^{(w)} - \nu_i^{(\xi)}}, \tag{16c}$$

$$\min_{(i,m) \in \mathcal{M}^{[<]}, \, \varepsilon_{i,m}^\top \nu^{(w)} > \nu_i^{(\eta)}} -\frac{\varepsilon_{i,m}^\top w - \eta_i + \varepsilon_{i,m}^\top \mu^{(w)} - \mu_j^{(\eta)}}{\varepsilon_{i,m}^\top \nu^{(w)} - \nu_i^{(\eta)}}, \tag{16d}$$

$$\min_{(i,m) \in \mathcal{M}^{[>]}, \, \varepsilon_{i,m}^\top \nu^{(w)} < \nu_i^{(\eta)}} -\frac{\varepsilon_{i,m}^\top w - \eta_i + \varepsilon_{i,m}^\top \mu^{(w)} - \mu_j^{(\eta)}}{\varepsilon_{i,m}^\top \nu^{(w)} - \nu_i^{(\eta)}}, \tag{16e}$$

$$\min_{j \in \mathcal{Z}, \nu_j^{(\alpha)} > 0} -\frac{\mu_j^{(\alpha)}}{\nu_j^{(\alpha)}}, \tag{16f}$$

$$\min_{(i,h) \in \mathcal{H}^{[=]}, \nu_{(i,h)}^{(\beta)} < 0, |\mathcal{H}_i^{[<]}| \leq \kappa - 2, |\mathcal{H}_i^{[=]}| \geq 2} -\frac{\mu_{(i,h)}^{(\beta)}}{\nu_{(i,h)}^{(\beta)}}, \tag{16g}$$

$$\min_{(i,h) \in \mathcal{H}^{[=]}, \nu_{(i,h)}^{(\beta)} > 0, |\mathcal{H}_i^{[<]}| < |\mathbb{H}_i| - \kappa, |\mathcal{H}_i^{[=]}| \geq 2} -\frac{\mu_{(i,h)}^{(\beta)}}{\nu_{(i,h)}^{(\beta)}}, \tag{16h}$$

$$\min_{(i,m) \in \mathcal{M}^{[=]}, \nu_{(i,m)}^{(\gamma)} < 0, |\mathcal{M}_i^{[<]}| \leq \kappa - 2, |\mathcal{M}_i^{[=]}| \geq 2} -\frac{\mu_{(i,m)}^{(\gamma)}}{\nu_{(i,m)}^{(\gamma)}}, \tag{16i}$$

$$\min_{(i,m) \in \mathcal{M}^{[=]}, \nu_{(i,m)}^{(\gamma)} > 0, |\mathcal{M}_i^{[<]}| < |\mathbb{M}_i| - \kappa, |\mathcal{M}_i^{[=]}| \geq 2} -\frac{\mu_{(i,m)}^{(\gamma)}}{\nu_{(i,m)}^{(\gamma)}} \Big\}, \tag{16j}$$

where $\min_{\theta \geq \theta_t}(z) := \max\{\theta_t, \min z\}$. In addition, if there are no index or pair of indices that applies to the conditions in (16), we define $\theta_{t+1} := \infty$.

Finally, we need to note a spacial case that the 2nd to 5th rules in Lemma 6 are applied during path-tracking. To explain such case, let us consider a situation that we have $(i, h) \in \mathcal{H}^{[=]}$ such that $\beta_{(i,h)} < 0$, $|\mathcal{H}_i^{[<]}| = \kappa - 1$ and $|\mathcal{H}_i^{[=]}| = 1$ at a breakpoint θ_{t-1}. In this case, we could not apply the 2nd rule in Lemma 6 to (i, h) because the conditions on the sizes of $\mathcal{H}_i^{[<]}$ and $\mathcal{H}_i^{[=]}$ are not satisfied. However, if one of the $(i, h') \in \mathcal{H}^{[<]}$ moves from $\mathcal{H}_i^{[<]}$ to $\mathcal{H}_i^{[=]}$ at the next breakpoint θ_t (and if $\beta_{(i,h)}$ is still negative), we could immediately move (i, h) from $\mathcal{H}_i^{[=]}$ to $\mathcal{H}_i^{[<]}$ by applying the 2nd rule. This situation indicates that there might be two simultaneous active set changes at a breakpoint. (In the pseudo-code below, we set $\theta_{t+1} = \theta_t$ in such situation.)

The regularization path tracking algorithm is summarized in Algorithm 2.

G   An Illustrative Example of The Algorithm Behavior

Let us first illustrate the behavior of the proposed algorithm using a simple toy example. Consider a binary classification problem with sample size $n = 7$ and the number of features $\ell = 2$. Figure 3(a) shows the scatter plot of the data set in the Euclidean space where each instance is represented as ⓪, ①, ②, ③, 4, 5 and 6. The instances ⓪, ①, ② and ③ belong to the first class (red circle), while the instances 4, 5 and 6 belong to the second class (blue square).

Algorithm 2 Regularization Path Tracking Algorithm for NN Feature Weighting

1: **Inputs**: A local optimal solution w at θ_0;
2: Initialize (ξ, η) and $(\mathcal{H}, \mathcal{M}, \mathcal{Z}, \mathcal{P})$ based on w and set $t \leftarrow 0$;
3: **for** $t = 0, 1, 2, \cdots$ **do**
4:     Compute the Lagrange multipliers β and γ;
5:     **if** Any rules in Lemma 6 applies to any active constraints in $\mathcal{H}^{[=]}$ or $\mathcal{M}^{[=]}$ **then**
6:         Update $\{\mathcal{H}, \mathcal{M}\}$ according to the rule, and set $\theta_{t+1} \leftarrow \theta_t$ and $t \leftarrow t + 1$;
7:     **else**
8:         Compute $\mu^{(w,)} \ldots, \mu^{(\gamma)}$ and $\nu^{(w)} \ldots, \nu^{(\gamma)}$ from (11), (12) and (13);
9:         Compute the next breakpoint θ_{t+1} from (16) and update $(\mathcal{H}, \mathcal{M}, \mathcal{Z}, \mathcal{P})$;
10:        **if** $\theta_{t+1} < \infty$ **then**
11:            $t \leftarrow t + 1$;
12:        **else**
13:            Terminate the regularization path;
14:   **end if**
15:     **end if**
16: **end for**
17: **Outputs**: The regularization path for $\theta \geq \theta_0$;

We applied the regularization path tracking to this toy data from the Euclidean feature space ($\bar{w} = \ell^{-1}\mathbf{1}$) with $\kappa = \lambda = 2$, i.e., the goal is to maximize the average $(2,2)$-neighbor margins. The SQP-FW algorithm with $\kappa = \lambda = 2$ behaves similarly. The algorithm experienced 18 active set changes and 17 of them were TN changes. Figure 2(a) shows the sequence of the weights w_1 and w_2 as the function of the 18 algorithm steps. The algorithm encountered TN changes in the first 17 steps, and observed an active-set change from $\mathcal{P}$ to $\mathcal{Z}$ at the end (the weight in the 2nd feature w_2 vanishes). Figure 2(b) shows the sequence of the weights w_1 and w_2 as functions of the regularization parameter θ. At $\theta = 0.606$ and 0.896, multiple TN changes were observed.
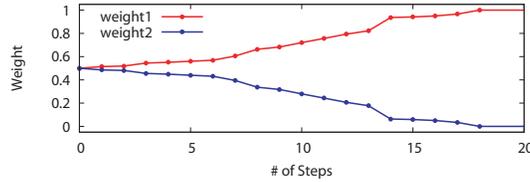
(a) The path of the weights with respect to the 18 algorithm steps
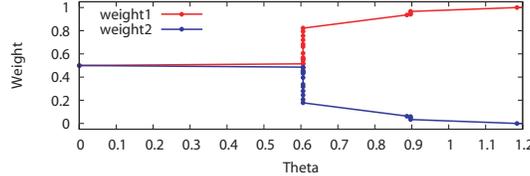
(b) The path of the weights with respect to the regularization parameter θ

Figure 2: The entire solution path of the weights in toy example.

The panel (a) in Figure 3 shows the initial Euclidean feature space, and the panels (b)–(d) show the optimally weighted feature space for several values of θ. Note that the horizontal and vertical axes in each panel are scaled according to the ratio $\sqrt{w_1} : \sqrt{w_2}$ in each panel. This is because the weighted distance between instance i and i' is represented as $\sqrt{w_1(x_{i,1} - x_{i',1})^2 + w_2(x_{i,2} - x_{i',2})^2}$. Panels (b) and (c) are the snapshots at two TN-change points at $\theta = 0.606$ and 0.887, respectively. The panel (d) shows the situation where only the first feature remains at the end of the algorithm.

Table 3 displays the sequences of TNs in each step of the algorithm. In this toy data set, feature 1 is more important than feature 2 for classification, so the algorithm monotonically increased w_1 and decreased w_2. The sequences in Table 3 show how the 2nd target hits and misses were switched

16

(a) Initial Euclidean feature space at $\theta = 0$. $w_1 = w_2 = 0.500$ are the solution.

(b) Weighted feature space at $\theta = 0.606$. $w_1 = 0.720$ and $w_2 = 0.280$ are the solution. A TN-change happens and the 2nd target miss of instance ⓪ is replaced from ④ to ⑥ .

(c) Weighted feature space at $\theta = 0.887$. $w_1 = 0.941$ and $w_2 = 0.059$ are the solution. A TN-change happens and the 2nd target miss of instance ⑥ is replaced from ① to ⓪.

(d) Weighted feature space at $\theta \geq 1.185$. $w_1 = 1$ and $w_2 = 0$ are the solution. Only the 1st feature is selected.

Figure 3: A simple illustrative example of the algorithm for a binary classification problem with sample size $n = 7$ and the number of features $\ell = 2$. The regularization path tracking algorithm was applied to this training set from the Euclidean feature space ($\bar{w} = \ell^{-1}\mathbf{1}$) with $\kappa = \lambda = 2$, i.e., maximizing the average $(2, 2)$-neighbor margins. (The SQP-FW algorithm with $\kappa = \lambda = 2$ behaves similarly.) Panel (a) is the initial Euclidean feature space at $\theta = 0$. Until the algorithm terminates at $\theta = 1.185$, 17 TN-changes were found. Panels (b) and (c) show the snapshots of the weighted feature space at two of them. Panel (d) is the weighted feature space at $\theta \geq 1.185$, where only the first feature was selected.

when the algorithm modified the 2-dimensional feature space from the Euclidean metric (a) to the low-rank metric (d).

Table 3: The sequence of the 2nd target hit and miss in toy example.

i	The sequence of the 2nd target hit h_i^2
0	2→2→1→1→1→3→3→3→3→3→3→3→3→3→3→3→3→3→3→3
1	0→0→0→3→3→3→3→3→3→3→3→3→3→3→3→3→3→3→2→2
2	1→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1
3	0→0→0→0→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1
4	6→6→6→6→6→6→6→6→6→6→6→6→6→6→6→5→5→5
5	4→4→4→4→4→4→4→4→4→4→4→4→4→4→4→4→4→4
6	4→4→4→4→4→4→4→4→4→4→4→4→4→4→4→4→4→4

i	The sequence of the 2nd target miss m_i^2
0	5→5→5→5→5→5→5→4→4→4→6→6→6→6→6→6→6→6→6
1	5→5→5→5→5→5→5→5→5→5→5→5→4→4→6→6→6→6→6
2	4→4→4→4→4→4→6→6→6→6→6→6→6→6→6→6→6→6
3	5→5→5→5→5→5→5→5→6→6→6→6→6→6→6→6→6→6
4	0→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1→1
5	2→2→2→2→2→2→2→2→1→1→1→0→0→0→0→0→0→0→0
6	3→3→3→3→3→3→3→3→3→3→3→3→3→1→1→0→0→0→0

17