A. Proof of Theorem 6

Theorem 6. *Suppose* $\mathbb{E}[N_i|\mathcal{H}_{t_i}] = \int_0^{t_i} g^*(w^* \cdot x_t)dt$, *where* g^* *is monotonic increasing, 1-Lipschitz and* $\|w^*\| \leq W$. *Then with probability at least* $1-\delta$, *there exist some iteration* $k < O\Big(\big(\frac{Wn}{\log(Wn/\delta)}\big)^{1/3}\Big)$ *such that*

$$\varepsilon(\hat{g}^k, \hat{w}^k) \leq O\bigg(\Big(\frac{W^2 \log(Wn/\delta)}{n}\Big)^{1/3}\bigg).$$

Notations. We define some extra notations. First we rewrite the integral as $\int_0^{t_i} g^*(w^* \cdot x_t)dt = \sum_{j\in\mathcal{S}_i} a_{ij} g^*(w^* \cdot x_j)$. Set $y_i^* = g^*(w^* \cdot x_i)$ to be the expected value of each y_i. Let $\bar{N}_i$ be the expected value of N_i. Then we have $\bar{N}_i = \sum_{j\in\mathcal{S}_i} a_{ij} y_j^*$. Clearly we do not have access to $\bar{N}_i$. However, consider a hypothetical call to the algorithm with input $\{(x_i, \bar{N}_i)\}_{i=1}^n$ and suppose it returns $\bar{g}^k$. In this case, we define $\bar{y}_i^k = \bar{g}^k(\bar{w}^k \cdot x_i)$. Next we begin the proof and introduce Lemma 3-5.

Analysis roadmap. To prove Theorem 6, we establish several lemmas. The heart of the proof is Lemma 3, in which we show a property of the learned parameters $\hat{w}^k$ at iteration k. That is, the squared distance $\|\hat{w}^k - w^*\|^2$ between $\hat{w}^k$ and the true direction w^* decreases at each iteration at a rate which depends on $\varepsilon(\hat{g}^k, \hat{w}^k)$ and some other additive error terms η_1 and η_2, which can be bounded respectively:

$$\|\hat{w}^k - w^*\|^2 - \|\hat{w}^{k+1} - w^*\|^2 \geq C_2 \varepsilon(\hat{g}^k, \hat{w}^k) - C_1(\eta_1 + \eta_2) \tag{16}$$

Lemma 4 bounds $\eta_1 = O\Big((K + \sqrt{4K^2 + 8k^2})\big(\log(\frac{1}{\delta})\big)^{1/2}\Big)$ using martingale concentration inequality.

Lemma 5 bounds $\eta_2 = O\bigg(\Big(\frac{W^2 \log(Wn/\delta)}{n}\Big)^{1/3}\bigg)$. It relates $\hat{y}_j^k$ (the value we can actually compute) and $\bar{y}_j^k$ (the value we could compute if we had $\bar{N}_i$). $\bar{y}_j^k$ and $\hat{y}_j^k$ will show up when we decouple $\|\hat{w}^k - w^*\|^2 - \|\hat{w}^{k+1} - w^*\|^2$.

Finally, we plug in the values of η_1 and η_2 to Lemma 3. Then we conduct telescoping sum of (16) and show there is at most $O\Big(W/(\eta_1 + \eta_2)\Big)$ iterations before the error $\varepsilon(\hat{g}^k, \hat{w}^k)$ is less than $O(\eta_1 + \eta_2)$. Since η_2 is the dominant term compared with η_1, we replace η_1 by η_2 in the final results. This completes the proof.

Now we introduce Lemma 3-5 as follows.

Lemma 3. *Suppose that* $\|w^k - w\| \leq W$, $\|x_i\| \leq 1$, $\sqrt{c} \leq \sum_{j\in\mathcal{S}_i} a_{ij} \leq \sqrt{C}, \forall i \in [n], j \in [n]$ *and* $y_j \leq M, \forall j \in [n]$, *and*

$$|\frac{1}{n}\sum_{i=1}^n (N_i - \bar{N}_i)| \leq \eta_1, \quad \frac{1}{n}\sum_{i=1}^n \sum_{j\in\mathcal{S}_i} a_{ij} |\hat{y}_j^k - \bar{y}_j^k| \leq \eta_2$$

then the following formula holds:

$$\|\hat{w}^k - w^*\|^2 - \|\hat{w}^{k+1} - w^*\|^2 \geq C_2 \varepsilon(\hat{g}^k, \hat{w}^k) - C_1(\eta_1 + \eta_2) \tag{17}$$

where $C_1 = \max\{5CW, 4M\sqrt{c} + 2CW\}$, $C_2 = 2c - C$.

The complete proof of Lemma 3 is in Appendix C.

Lemma 4 (Martingale Concentration Inequality)**.** *Suppose* $dM(t) \leq K$, $V(t) \leq k$ *for all* $t > 0$ *and some* $K, k \geq 0$. *With probability at least* $1-\delta$, *it holds that*

$$\frac{1}{n}\sum_{i=1}^n |N_i - \bar{N}_i| \leq O\left((K + \sqrt{4K^2 + 8k^2})\big(\log(1/\delta)\big)^{1/2}\right).$$

Note $N_i - \bar{N}_i = M_i$, which is the martingale at time t_i. A continuous martingale is a stochastic process such that $\mathbb{E}[M_t|\{M_\tau, \tau \leq s\}] = M_s$. It means the conditional expectation of an observation at time t is equal to the observation at time s, given all the observations up to time $s \leq t$. $V(t)$ is the variation process. It is shown in (Aalen et al., 2008) that $V(t) = \Lambda(t) = \int_0^t \lambda(s)ds$, which is the compensator for point process $N(t)$. The martingale serves as the noise term in point processes (similar to Gaussian noise in regression) and can be bounded using the Bernstein-type concentration inequality. The proof is in Appendix D.

Lemma 5. *(Kakade et al., 2011) With probability at least $1-\delta$, it holds for any k that*

$$\frac{1}{n}\sum_{j=1}^{n}|\hat{y}_j^k - \bar{y}_j^k| \leq O\bigg(\Big(\frac{W^2\log(Wn/\delta)}{n}\Big)^{1/3}\bigg).$$

Lemma 5 relates $\hat{y}_j^k$ (the value we can actually compute) to $\bar{y}_j^k$ (the value we could compute if we had the conditional means of N_j). The proof of this lemma uses the covering number technique and can be found in (Kakade et al., 2011).

Proof of Theorem 6. With Lemma 3, we can conduct telescoping sum. There can be two cases: either $\varepsilon(\hat{g}^k, \hat{w}^k) \leq 3C_1(\eta_1+\eta_2)/C_2$ or $\varepsilon(\hat{g}^k, \hat{w}^k) \geq 3C_1(\eta_1+\eta_2)/C_2$. If it is the first case, then we are done. If it is the second case, then we have:

$$\|w^k - w\|^2 - \|w^{k+1} - w\|^2 \geq C_1(\eta_1+\eta_2)$$

Since $\|w^{k+1}-w\|^2 \geq 0$, and $\|w^0 - w\|^2 \leq 2W^2$, by telescoping sum, at iteration K, we have:

$$2W^2 \geq \|w^0 - w\|^2 - \|w^K - w\|^2 \geq KC_1(\eta_1+\eta_2)$$

Set $K = 2W^2/C_1(\eta_1+\eta_2)$, if $k > K$, then the above inequality does not hold, which means $\varepsilon(\hat{g}^k, \hat{w}^k) \geq 3C_1(\eta_1+\eta_2)/C_2$ does not hold. Hence there can be at most $2W^2/C_1(\eta_1+\eta_2) = O(W/(\eta_1+\eta_2))$ iterations before $\varepsilon(\hat{g}^k, \hat{w}^k) \leq 3C_1(\eta_1+\eta_2)/C_2$.

The remaining step is to bound η_1 and η_2. We use Lemma 4 to bound η_1 and use Lemma 5 to bound η_2. Clearly η_2 is the dominant term. Plugging the values of η_1 and η_2, we have the conclusion that there is some h^k such that

$$\varepsilon(\hat{g}^k, \hat{w}^k) \leq O\bigg(\Big(\frac{W^2\log(Wn/\delta)}{n}\Big)^{1/3}\bigg)$$

B. Proof of Lemma 6

To prove Lemma 3, a key technique is the generalized calibration property. It generalizes that of isotonic regression in (Kalai & Sastry, 2009) since our objective function is more general. We first state Lemma 6 and then provide the proof.

Lemma 6 (Generalized Calibration Property)**.** *The solutions to Quadratic Problem in (11) is partitioned into disjoint blocks* $\{\mathcal{P}_l\}_{l=1}^m$*, and for each block* $\mathcal{P}_l$*:*

$$\sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k)\sum_{j\in\mathcal{P}_l} a_{ij} = 0 \tag{18}$$

Proof. First we define a_{ij} such that

$$a_{ij} = \begin{cases} a_{ij} & \text{if } j \in \mathcal{S}_i \\ 0 & \text{else} \end{cases}$$

Hence we have

$$\sum_{j=1}^{n} a_{ij} = \sum_{j\in\mathcal{S}_i} a_{ij} \tag{19}$$

We can rewrite the objective function as:

$$f = \frac{1}{2}\sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k)^2 = \frac{1}{2}\sum_{i=1}^{n}(N_i - \sum_{j=1}^{n} a_{ij}\hat{y}_j^k)^2$$

Set $\{\lambda_i\}_{i=1}^{n-1}$ to be the Lagrange multipliers. To update $\hat{y}_j^k$, we apply the KKT conditions to (11) and obtain the following formulas:

$$\frac{\partial f}{\partial \hat{y}_1^k} = \sum_{i=1}^{n}(N_i - \sum_{j=1}^{n} a_{ij}\hat{y}_j^k)a_{i1} + \lambda_1 = 0 \tag{20}$$

$$\frac{\partial f}{\partial \hat{y}_j^k} = \sum_{i=1}^{n}(N_i - \sum_{j=1}^{n} a_{ij}\hat{y}_j^k)a_{ij} + \lambda_j - \lambda_{j-1} = 0, \quad 2 \le j \le n-1 \tag{21}$$

$$\frac{\partial f}{\partial \hat{y}_n^k} = \sum_{i=1}^{n}(N_i - \sum_{j=1}^{n} a_{ij}\hat{y}_j^k)a_{in} - \lambda_{n-1} = 0 \tag{22}$$

$$\lambda_j(\hat{y}_j^k - \hat{y}_{j+1}^k) = 0, \quad 1 \le j \le n-1 \tag{23}$$

$$\lambda_j \ge 0, \quad 1 \le j \le n-1 \tag{24}$$

Depending whether $\hat{y}_j^k$'s are equal, we can divide the subscript of $\hat{y}_j^k$ into disjoint sets $\{\mathcal{P}_l\}_{l=1}^m$ such that in each $\mathcal{P}_l$, the values of $\hat{y}_j^k$ are the same. Hence there exists $j_1 < j_2 < \cdots < j_{m-1} < n$, such that

$$\mathcal{P}_1 = \{1, \cdots, j_1\}, \mathcal{P}_2 = \{j_1+1, \cdots, j_2\}, \cdots, \mathcal{P}_m = \{j_{m-1}+1, n\} \tag{25}$$

Figure 8 illustrates an example when $m = 3$. in this case, $\mathcal{P}_1 = \{1, 2\}$, $\mathcal{P}_2 = \{3, 4\}$, and $\mathcal{P}_3 = \{5, 6\}$. Now we show the following equality holds for $l = 1, \cdots, m$ in three cases,

$$\sum_{i=1}^{n}(N_i - \sum_{j=1}^{n} a_{ij}\hat{y}_j^k)\sum_{j\in\mathcal{P}_l} a_{ij} = 0$$

Case 1: the first block. For $\mathcal{P}_1$, we sum up equations $\frac{\partial f}{\partial \hat{y}_j^k} = 0$ according to the index in $\mathcal{P}_1$. we have

$$\begin{cases} \sum_{i=1}^{n}(N_i - \sum_{j=1}^{n} a_{ij}\hat{y}_j^k)\sum_{j\in\mathcal{P}_1} a_{ij} + \lambda_{j_1} = 0 \\ \lambda_{j_1} = 0 \end{cases} \tag{26}$$

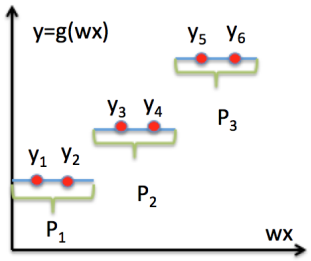

Figure 8. Demonstration for the block partition in (25). $\mathcal{P}_1$, $\mathcal{P}_2$ and $\mathcal{P}_3$ are the first, intermediate and last block respectively. In each block, $\hat{y}$ has the same value.

Since $\hat{y}^k_{j_1} \neq \hat{y}^k_{j_1+1}$, from (23) we have $\lambda_{j_1} = 0$.

Case 2: the intermediate blocks. For $2 \leq l \leq m-1$, in $\mathcal{P}_l$, we sum up equations $\frac{\partial f}{\partial \hat{y}^k_j} = 0$. Then we have

$$\begin{cases} \sum_{i=1}^n (N_i - \sum_{j=1}^n a_{ij}\hat{y}^k_j) \sum_{j\in\mathcal{P}_l} a_{ij} + \lambda_{j_l} - \lambda_{j_{l-1}} = 0 \\ \lambda_{j_l} = \lambda_{j_{l-1}} = 0 \end{cases} \tag{27}$$

Since $\hat{y}^k_{j_l} \neq \hat{y}^k_{j_l+1}$ and $\hat{y}^k_{j_{l-1}} \neq \hat{y}^k_{j_{l-1}+1}$, from (23) we have $\lambda_{j_l} = \lambda_{j_{l-1}} = 0$.

Case 3: the last block. For $\mathcal{P}_m$, similarly we have

$$\begin{cases} \sum_{i=1}^n (N_i - \sum_{j=1}^n a_{ij}\hat{y}^k_j) \sum_{j\in\mathcal{P}_m} a_{ij} - \lambda_{j_{m-1}} = 0 \\ \lambda_{j_{m-1}} = 0 \end{cases} \tag{28}$$

From (19), we have for all $l = 1, \cdots, m$

$$\sum_{i=1}^n (N_i - \sum_{j\in\mathcal{P}_l} a_{ij}\hat{y}^k_j) \sum_{j\in\mathcal{P}_l} a_{ij} = 0$$

This completes the proof.

C. Proof of Lemma 3

First, we have

$$\|\hat{w}^k - w^*\|^2 - \|\hat{w}^{k+1} - w^*\|^2 = 2(\hat{w}^{k+1} - \hat{w}^k) \cdot (w^* - \hat{w}^k) - \|\hat{w}^{k+1} - \hat{w}^k\|^2 \tag{29}$$

$$= \underbrace{\frac{2}{n}\sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k)(\sum_{j\in\mathcal{S}_i} a_{ij}x_j \cdot (w^* - \hat{w}^k))}_{A} - \underbrace{\left\|\frac{1}{n}\sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k)\sum_{j\in\mathcal{S}_i} a_{ij}x_j\right\|^2}_{B} \tag{30}$$

First we simplify A. Using the following equality:

$$N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k = N_i - \sum_{j\in\mathcal{S}_i} a_{ij}y_j^* + \sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k + \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k,$$

we can rewrite A into three parts:

$$A = \frac{2}{n}\sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}y_j^*)(\sum_{j\in\mathcal{S}_i} a_{ij}x_j) \cdot (w^* - w^k) \tag{31}$$

$$+ \frac{2}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)(\sum_{j\in\mathcal{S}_i} a_{ij}x_j \cdot (w^* - w^k)) \tag{32}$$

$$+ \frac{2}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k)(\sum_{j\in\mathcal{S}_i} a_{ij}x_j \cdot (w^* - w^k)) \tag{33}$$

The term (31) is at least $-2CW\eta_1$, the term (33) is at least $-2CW\eta_2$ since $|\sum_{j\in\mathcal{S}_i} a_{ij}(w - w^k) \cdot x_j| \leq \sqrt{C}W$ and assuming $C \geq 1$. We thus bound(32).

First define v, the inverse of g as

$$v(y) = \inf\{z \in \text{dom}(g) | g(z) = y\}$$

Note that v is well defined since g is monotonic. We also split (32) into three parts,

$$\frac{2}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)(\sum_{j\in\mathcal{S}_i} a_{ij}x_j \cdot (w^* - \hat{w}^k))$$

$$= \frac{2}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)\sum_{j\in\mathcal{S}_i} a_{ij}v(\bar{y}_j^k) \tag{34}$$

$$- \frac{2}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)\sum_{j\in\mathcal{S}_i} a_{ij}\hat{w}^k \cdot x_j \tag{35}$$

$$+ \frac{2}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)\sum_{j\in\mathcal{S}_i} a_{ij}(w^* \cdot x_j - v(\bar{y}_j^k)) \tag{36}$$

As to (34), it is 0 by Lemma 6. To see this, remember that $\bar{N}_i = \sum_{j\in\mathcal{S}_i} a_{ij}y_i^*$ and $\bar{y}_j^k$ is the output of the algorithm in Eq. (11) with input $\{(\bar{w}^k \cdot x_i, \bar{N}_i)\}$. Apply Lemma 6 and we have the pools $\{\mathcal{P}_l\}_{l=1}^m$ and

$$\sum_{i=1}^{n}(\bar{N}_i - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)\sum_{j\in\mathcal{P}_l}\sum_{j\in\mathcal{S}_i} a_{ij} = 0$$

Define function v to be the inverse of g. v is defined as $v(y) = \inf\{z \in \text{dom}(g) | g(z) = y\}$. Since g is monotonic, v is well-defined. Since all $\bar{y}_j^k$ in the same set $\mathcal{P}_l$ has the same value, then the value $v(\bar{y}_j^k)$ (the inverse mapping) is also the same. Hence

$$\sum_{i=1}^{n}(N_i - \sum_{j\in\mathcal{S}_i} a_{ij}\bar{y}_j^k)\sum_{j\in\mathcal{S}_i} a_{ij}v(\bar{y}_j^k) = 0$$

Now sum the above equation up for all sets $\mathcal{P}_l, l = 1, \cdots, m$, note that $\bigcup_{l=1}^{m} \mathcal{P} = \{1, \cdots, n\}$, we have

$$\sum_{i=1}^{n} (\bar{N}_i - \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k) \sum_{j \in \mathcal{S}_i} a_{ij} v(\bar{y}_j^k) = 0$$

As to (35), we show it is always no greater than 0. To see this, we first claim that for any $\delta > 0$,

$$\sum_{i=1}^{n} (\bar{N}_i - \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k)^2 \leq \sum_{i=1}^{n} (\bar{N}_i - \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k - \delta (\sum_{j \in \mathcal{S}_i} a_{ij} x_j) \cdot \hat{w}^k)^2$$

This is because $\sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k$ minimizes the sum of squared difference w.r.t. $\bar{N}_i$ over all such sequences. Rewriting this as a difference of squares gives,

$$\sum_i \delta (\sum_{j \in \mathcal{S}_i} a_{ij} x_j) \cdot \hat{w}^k \left(2N_i - 2 \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k - \delta (\sum_{j \in \mathcal{S}_i} a_{ij} x_j) \cdot \hat{w}^k \right) \geq 0$$

Dividing both sides by $2\delta > 0$, we have

$$\sum_i (\sum_{j \in \mathcal{S}_i} a_{ij} x_j) \cdot \hat{w}^k \left(\bar{N}_i - \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k - \frac{\delta}{2} (\sum_{j \in \mathcal{S}_i} a_{ij} x_j) \cdot \hat{w}^k \right) \geq 0$$

Setting $\delta \to 0$, by continuity we obtain

$$\frac{2}{n} \sum_{i=1}^{n} (\bar{N}_i - \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k) \sum_{j \in \mathcal{S}_i} a_{ij} \hat{w}^k \cdot x_j \geq 0$$

Hence we have (35) always no greater than 0.

As to (36), by 1-Lipschitz property of g, the first term can be bounded as

$$\begin{aligned}
&\frac{2}{n} \sum_{i=1}^{n} (\sum_{j \in \mathcal{S}_i} a_{ij} y_j^* - \sum_{j \in \mathcal{S}_i} a_{ij} \bar{y}_j^k) \sum_{j \in \mathcal{S}_i} a_{ij} (v(y_j^*) - v(\bar{y}_j^k)) \\
&\geq \frac{2}{n} \sum_{j=1}^{n} c(y_j^* - \bar{y}_j^k)(v(y_j^*) - v(\bar{y}_j^k)) \\
&\geq \frac{2}{n} \sum_{j=1}^{n} c(y_j^* - \bar{y}_j^k)^2 = 2c\varepsilon(\bar{g}^k, \bar{w}^k) \qquad (37)
\end{aligned}$$

Plugging to the definition of A, we get

$$\boxed{A \geq 2c\varepsilon(\bar{g}^k, \bar{w}^k) - 2CW(\eta_1 + \eta_2)} \qquad (38)$$

Next we bound B. First rewrite B as:

$$\begin{aligned}
B &= \left\| \frac{1}{n} \sum_{i=1}^{n} (N_i - \sum_{j \in \mathcal{S}_i} a_{ij} y_j^* + \sum_{j \in \mathcal{S}_i} a_{ij} y_j^* - \sum_{j \in \mathcal{S}_i} a_{ij} \hat{y}_j^k) \sum_{j \in \mathcal{S}_i} a_{ij} x_j \right\|^2 \\
&\leq \left\| \frac{1}{n} \sum_{i=1}^{n} (N_i - \sum_{j \in \mathcal{S}_i} a_{ij} y_j^*) \sum_{j \in \mathcal{S}_i} a_{ij} x_j \right\|^2 \qquad (39) \\
&+ 2 \left\| \frac{1}{n} \sum_{i=1}^{n} (N_i - \sum_{j \in \mathcal{S}_i} a_{ij} y_j^*) \sum_{j \in \mathcal{S}_i} a_{ij} x_i \right\| \times \left\| \frac{1}{n} \sum_{i=1}^{n} (\sum_{j \in \mathcal{S}_i} a_{ij} y_j^* - \sum_{j \in \mathcal{S}_i} a_{ij} \hat{y}_j^k) \sum_{j \in \mathcal{S}_i} a_{ij} x_j \right\| \qquad (40) \\
&+ \left\| \frac{1}{n} \sum_{i=1}^{n} (\sum_{j \in \mathcal{S}_i} a_{ij} y_j^* - \sum_{j \in \mathcal{S}_i} a_{ij} \hat{y}_j^k) \sum_{j \in \mathcal{S}_i} a_{ij} x_j \right\|^2 \qquad (41)
\end{aligned}$$

From the condition in Lemma 3, we have

$$\|\frac{1}{n}\sum_{i=1}^{n}(N_i - \frac{1}{n}\sum_{i=1}^{n}\sum_{j\in\mathcal{S}_i} a_{ij}y_j^*)\sum_{j\in\mathcal{S}_i} a_{ij}x_j\|^2 \leq C\eta_1^2 \tag{42}$$

Use Jensen's inequality and consider the upper bound C for $\|\sum_{j\in\mathcal{S}_i} a_{ij}x_i\|^2$, we show that

$$\left\|\frac{1}{n}\sum_{i=1}^{n}(\sum_{j\in\mathcal{S}_i} a_{ij}y_j^* - \sum_{j\in\mathcal{S}_i} a_{ij}\hat{y}_j^k)\sum_{j\in\mathcal{S}_i} a_{ij}x_i\right\|^2 \leq C\times\frac{1}{n}\sum_{i=1}^{n}(y_j^* - \hat{y}_j^k)^2 = C\varepsilon(\hat{g}^k,\hat{w}^k) \tag{43}$$

Combining (42) and (43) into (39), (40), (41), assuming $\eta_1 \leq 1$, $C \geq 1$, we have

$$B \leq C\eta_1^2 + 2C\eta_1\sqrt{\varepsilon(\hat{g}^k,\hat{w}^k)} + C\varepsilon(\hat{g}^k,\hat{w}^k) \leq C\varepsilon(\hat{g}^k,\hat{w}^k) + 3C\eta_1 \tag{44}$$

Hence the we have:

$$\boxed{B \leq C\varepsilon(\hat{g}^k,\hat{w}^k) + 3C\eta_1} \tag{45}$$

Combining the bound for A in (38) and the bound for B in (45) into (30), we get

$$\boxed{\|\hat{w}^k - \hat{w}\|^2 - \|\hat{w}^{k+1} - \hat{w}\|^2 \geq 2c\varepsilon(\bar{g}^k,\bar{w}^k) - C\varepsilon(\hat{g}^k,\hat{w}^k) - CW(5\eta_1 + 2\eta_2)} \tag{46}$$

To finish the proof, we establish the relationship between $\varepsilon(\bar{g}^k,\bar{w}^k)$ and $\varepsilon(\hat{g}^k,\hat{w}^k)$ as follows: we claim that the difference between $\varepsilon(\bar{g}^k,\bar{w}^k)$ and $\varepsilon(\hat{g}^k,\hat{w}^k)$ can be lower bounded:

$$\boxed{\varepsilon(\bar{g}^k,\bar{w}^k) - \varepsilon(\hat{g}^k,\hat{w}^k) \geq -2M\eta_2/\sqrt{c}} \tag{47}$$

To see this, we have:

$$\begin{aligned}
\varepsilon(\bar{g}^k,\bar{w}^k) &= \frac{1}{n}\sum_{j=1}^{n}(\bar{y}_j^k - y_j^*)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}(\bar{y}_j^k - \hat{y}_j^k + \hat{y}_j^k - y_j^*)^2 \\
&= \frac{1}{n}\sum_{j=1}^{n}(\hat{y}_j^k - y_j^*)^2 + \frac{1}{n}\sum_{j=1}^{n}(\bar{y}_j^k - \hat{y}_j^k)(\bar{y}_j^k + \hat{y}_j^k - 2y_j^*) \\
&= \varepsilon(\hat{g}^k,\hat{w}^k) + \frac{1}{n}\sum_{j=1}^{n}(\bar{y}_j^k - \hat{y}_j^k)(\bar{y}_j^k + \hat{y}_j^k - 2y_j^*)
\end{aligned}$$

and we have $|\bar{y}_i^k + \hat{y}_i^k - 2y_i^*| \leq 2M$. Plugging this and the following inequality leads to (47).

$$\frac{1}{n}\sum_{j=1}^{n}|\hat{y}_j^k - \bar{y}_j^k| \leq \frac{1}{n}\sum_{j=1}^{n}\sum_{j\in\mathcal{S}_i} a_{ij}/\sqrt{c}|\hat{y}_j^k - \bar{y}_j^k| \leq \eta_2/\sqrt{c}$$

Combine (47) and (46), we have

$$\boxed{\|w^k - w\|^2 - \|w^{k+1} - w\|^2 \geq (2c - C)\varepsilon(\hat{g}^k,\hat{w}^k) - 4M\sqrt{c}\eta_2 - CW(5\eta_1 + 2\eta_2) \geq C_2\varepsilon(\hat{g}^k,\hat{w}^k) - C_1(\eta_1 + \eta_2)}$$

where $C_1 = \max\{5CW, 4M\sqrt{c} + 2CW\}$, $C_2 = (2c - C)$, this completes the proof.

D. Proof of Lemma 4

We have $N_i - \bar{N}_i = M_i$, which is the martingale at time t_i. The martingale serves as the noise term in point processes (similar to Gaussian noise in regression) and can be bounded using the Bernstein-type concentration inequality. First, we have the following martingale inequality (Aalen et al., 2008; Liptser & Shiryayev, 2012): for each ϵ and some t, we have

$$\mathbb{P}[|M(t)| > \epsilon] \leq \exp\bigg(- \frac{\epsilon^2}{2(k^2 + \epsilon K)} \bigg)$$

In our case, for each i, we have $N_i = \Lambda(t_i) + M(t_i)$, where $\Lambda(t)$ is the compensator and $M(t)$ is the zero-mean martingale. Also we have $\bar{N}_i = \mathbb{E}(N_i) = \Lambda(t_i)$. Hence $N_i - \bar{N}_i = M(t_i) = M_i$. Now we set $\delta = \mathbb{P}[|M(t)| > \epsilon]$, then with probability at least $1 - \delta$, $|M(t)| \leq \epsilon$. Set $\delta = \exp\bigg(- \frac{\epsilon^2}{2(k^2 + \epsilon K)} \bigg)$, then we have the equation

$$\epsilon^2 - 2K \log(\frac{1}{\delta})\epsilon - 2k^2 \log(\frac{1}{\delta}) = 0$$

Hence

$$\begin{aligned}
\epsilon &= \frac{2K \log(\frac{1}{\delta}) + \sqrt{4K^2 \big(\log(\frac{1}{\delta})\big)^2 + 8k^2 \log(\frac{1}{\delta})}}{2} \\
&\leq K \log(\frac{1}{\delta}) + \sqrt{4K^2 + 8k^2} \bigg(\log(\frac{1}{\delta}) \bigg)^{1/2} \\
&\leq (K + \sqrt{4K^2 + 8k^2}) \bigg(\log(\frac{1}{\delta}) \bigg)^{1/2}
\end{aligned}$$

Here we have used the fact that $\big(\log(1/\delta)\big)^2 \leq \log(1/\delta) \leq \sqrt{\log(1/\delta)}$. We can obtain that

$$\epsilon = O\left((K + \sqrt{4K^2 + 8k^2}) \big(\log(\frac{1}{\delta}) \big)^{1/2} \right)$$

Hence we have

$$\frac{1}{n} \sum_{i=1}^{n} |N_i - \bar{N}_i| = \frac{1}{n} \sum_{i=1}^{n} |M_i| \leq O\left((K + \sqrt{4K^2 + 8k^2}) \big(\log(\frac{1}{\delta}) \big)^{1/2} \right)$$