Supplementary Material for Nonparametric Budgeted Stochastic Gradient Descent

1 Notion

We introduce some notions used in this supplementary material.

For regression task, we define $y_{\max} = \max_y |y|$. We further denote the set S as

$$S = \begin{cases} \mathcal{B}\left(\mathbf{0}, y_{\max}\lambda^{-1/2}\right) & \text{if L2 is used and } \lambda \leq 1 \\ \mathbb{R}^D & \text{otherwise} \end{cases}$$

where $\mathcal{B}\left(\mathbf{0}, y_{\max}\lambda^{-1/2}\right) = \left\{\mathbf{w} \in \mathbb{R}^D : \|\mathbf{w}\| \leq y_{\max}\lambda^{-1/2}\right\}$ and $\mathbb{R}^D$ specifies the whole feature space.

2 Loss Functions

We introduce five types of loss functions that can be used in our proposed algorithm, namely Hinge, Logistic, L2, L1, and $\varepsilon-$insensitive losses. We verify that these loss functions satisfying the necessary condition, that is, $\left\|l^{'}\left(\mathbf{w}; x, y\right)\right\| \leq A\left\|\mathbf{w}\right\|^{1/2} + B$ for some appropriate positive numbers A, B. Without loss of generality, we assume that feature domain are bounded, i.e., $\|\Phi(x)\| \leq 1$, $\forall x \in \mathcal{X}$.

- **Hinge loss**

$$l\left(\mathbf{w}; x, y\right) = \max\left\{0, 1 - y\mathbf{w}^{\mathsf{T}}\Phi(x)\right\}$$
$$l^{'}\left(\mathbf{w}; x, y\right) = -\mathbb{I}_{\{y\mathbf{w}^{\mathsf{T}}\Phi(x) \leq 1\}} y\Phi(x)$$

Therefore, by choosing $A = 0$, $B = 1$ we have

$$\left\|l^{'}\left(\mathbf{w}; x, y\right)\right\| = \|\Phi(x)\| \leq 1 = A\|\mathbf{w}\|^{1/2} + B$$

- **L2 loss**

In this case, at the outset we cannot verify that $\left\|l^{'}\left(\mathbf{w}; x, y\right)\right\| \leq A\|\mathbf{w}\|^{1/2} + B$ for all $\mathbf{w}, x, y$. However, to support the proposed theory, we only need to check that $\left\|l^{'}\left(\mathbf{w}_t; x, y\right)\right\| \leq A\|\mathbf{w}_t\|^{1/2} + B$ for all $t \geq 1$. We derive as follows

$$l\left(\mathbf{w}; x, y\right) = \frac{1}{2}\left(y - \mathbf{w}^{\mathsf{T}}\Phi(x)\right)^2$$
$$l^{'}\left(\mathbf{w}; x, y\right) = \left(\mathbf{w}^{\mathsf{T}}\Phi(x) - y\right)\Phi(x)$$

$$\begin{aligned}\left\|l^{'}\left(\mathbf{w}_t; x, y\right)\right\| &= \left|\mathbf{w}_t^{\mathsf{T}}\Phi(x) + y\right| \|\Phi(x)\| \leq \left|\mathbf{w}_t^{\mathsf{T}}\Phi(x)\right| + y_{\max} \\ &\leq \|\Phi(x)\| \|\mathbf{w}_t\| + y_{\max} \leq A\|\mathbf{w}_t\|^{1/2} + B\end{aligned}$$

where $B = y_{\max}$ and $A = \begin{cases} y_{\max}^{1/2}\lambda^{-1/4} & \text{if } \lambda \leq 1 \\ y_{\max}^{1/2}(\lambda - 1)^{-1/2} & \text{otherwise} \end{cases}$.

Here we note that we make use of the fact that $\|\mathbf{w}_t\| \leq y_{\max}(\lambda - 1)^{-1}$ if $\lambda > 1$ (cf. Thm. 7) and $\|\mathbf{w}_t\| \leq y_{\max}\lambda^{-1/2}$ otherwise (cf. Line 13 in Alg. 2 and Line 16 in Alg. 3).

- **L1 loss**

$$l\left(\mathbf{w};x,y\right)=|y-\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)|$$
$$l^{'}\left(\mathbf{w};x,y\right)=\text{sign}\left(\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)-y\right)\Phi\left(x\right)$$

Therefore, by choosing $A=0$, $B=1$ we have

$$\left\|l^{'}\left(\mathbf{w};x,y\right)\right\|=\left\|\Phi\left(x\right)\right\|\leq 1=A\left\|\mathbf{w}\right\|^{1/2}+B$$

- **Logistic loss**

$$l\left(\mathbf{w};x,y\right)=\log\left(1+\exp\left(-y\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)\right)\right)$$
$$l^{'}\left(\mathbf{w};x,y\right)=\frac{-y\exp\left(-y\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)\right)\Phi\left(x\right)}{\exp\left(-y\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)\right)+1}$$

Therefore, by choosing $A=0$, $B=1$ we have

$$\left\|l^{'}\left(\mathbf{w};x,y\right)\right\|<\left\|\Phi\left(x\right)\right\|\leq 1=A\left\|\mathbf{w}\right\|^{1/2}+B$$

- **ε-insensitive loss**

$$l\left(\mathbf{w};x,y\right)=\max\left\{0,|y-\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)|-\varepsilon\right\}$$
$$l^{'}\left(\mathbf{w};x,y\right)=\mathbb{I}_{\{|y-\mathbf{w}^{\mathsf{T}}\Phi(x)|>\varepsilon\}}\text{sign}\left(\mathbf{w}^{\mathsf{T}}\Phi\left(x\right)-y\right)x$$

Therefore, by choosing $A=0$, $B=1$ we have

$$\left\|l^{'}\left(\mathbf{w};x,y\right)\right\|=\left\|\Phi\left(x\right)\right\|\leq 1=A\left\|\mathbf{w}\right\|^{1/2}+B$$

3 Proofs

In this section, we present the full proofs of the corollaries and theorems in our paper.

Corollary 1. *The following holds for all t,*

$$\mathbb{E}\left[\left\|\mathbf{w}_t\right\|^2\right]<P^2=\left(\frac{A+\sqrt{A^2+B\lambda}}{\lambda}\right)^2$$

Proof. We prove by induction in t that $\mathbb{E}\left[\left\|\mathbf{w}_t\right\|^2\right]^{1/2}<P=\frac{A+\sqrt{A^2+B\lambda}}{\lambda}$, $\forall t=1,2,\ldots$

It is obvious for $t=1$ from $\mathbb{E}\left[\left\|\mathbf{w}_1\right\|^2\right]^{1/2}=0$.

Assume that the statement holds for t, according to Minkowski inequality we then have

$$\begin{aligned}\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_{t+1}\right\|^2\right]}\leq\frac{t-1}{t}\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_t\right\|^2\right]}+\frac{1}{\lambda t}\sqrt{\mathbb{E}\left[\left\|l'\left(\mathbf{w}_t;x_t,y_t\right)\right\|^2\right]}+\frac{1}{\lambda t}\sqrt{\mathbb{E}\left[Z_t^2\left\|l'\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right\|^2\right]}\\ \leq\frac{t-1}{t}\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_t\right\|^2\right]}+\frac{1}{\lambda t}\sqrt{\mathbb{E}\left[\left\|l'\left(\mathbf{w}_t;x_t,y_t\right)\right\|^2\right]}+\frac{1}{\lambda t}\sqrt{\mathbb{E}\left[\left\|l'\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right\|^2\right]}\\ \leq\frac{t-1}{t}\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_t\right\|^2\right]}+\frac{1}{\lambda t}\left(A\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_t\right\|\right]}+B+A\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_{t'}\right\|\right]}+B\right)\\ \leq\frac{t-1}{t}P+\frac{2}{\lambda t}\left(A\sqrt{P}+B\right)=P\end{aligned}$$

Note that we have used the assumption about loss function $\left\|l^{'}\left(\mathbf{w};x,y\right)\right\|\leq A\left\|\mathbf{w}\right\|^{1/2}+B$ for all $\mathbf{w},x,y$. □

Corollary 2. *The following holds for all t,*

$$\mathbb{E}\left[\left\|l^{'}\left(\mathbf{w}_t;x_t,y_t\right)\right\|^2\right]\leq L=\left(A\sqrt{P}+B\right)^2$$

Proof. We have the following

$$\sqrt{\mathbb{E}\left[\left\|l'\left(\mathbf{w}_t;x_t,y_t\right)\right\|^2\right]} \le \sqrt{\mathbb{E}\left[\left(A\left\|\mathbf{w}_t\right\|^{1/2}+B\right)^2\right]} \le A\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_t\right\|\right]}+B \le A\sqrt{P}+B$$

□

Corollary 3. *The following holds for all t,*

$$\mathbb{E}\left[\left\|g_t\right\|^2\right] \le G = \left(\lambda P + A\sqrt{P}+B\right)^2$$

Proof. Again using Minkowski inequality

$$\sqrt{\mathbb{E}\left[\left\|g_t\right\|^2\right]} \le \lambda\sqrt{\mathbb{E}\left[\left\|\mathbf{w}_t\right\|^2\right]} + \sqrt{\mathbb{E}\left[\left\|l'\left(\mathbf{w}_t;x_t,y_t\right)\right\|^2\right]} \le \lambda P + A\sqrt{P}+B$$

□

Corollary 4. *The following holds for all t,*

$$\mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2\right] \le W = \frac{\lambda L^{1/2}+\sqrt{\lambda^2 L + 8\lambda^2 Q}}{4\lambda^2}$$

Proof. Let us define $\delta_t = g_t - Z_t l^{'}\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)$. We have the following

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\delta_t$$

$$\begin{aligned}\left\|\mathbf{w}_{t+1}-\mathbf{w}^*\right\|^2 &= \left\|\mathbf{w}_t-\eta_t\delta_t-\mathbf{w}^*\right\|^2 = \left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2 + \eta_t^2\left\|\delta_t\right\|^2 - 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,\delta_t\right\rangle \\ &= \left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2 + \eta_t^2\left\|\delta_t\right\|^2 - 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,g_t\right\rangle + 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*, Z_t l^{'}\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right\rangle\end{aligned}$$

Taking conditional expectation w.r.t $\mathbf{w}_t^1$, we gain

$$\begin{aligned}\mathbb{E}\left[\left\|\mathbf{w}_{t+1}-\mathbf{w}^*\right\|^2\right] &= \mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2\right] + \eta_t^2\mathbb{E}\left[\left\|\delta_t\right\|^2\right] - 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,\mathbb{E}\left[g_t\right]\right\rangle + 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,\mathbb{E}\left[Z_t l^{'}\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right]\right\rangle \\ &= \mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2\right] + \eta_t^2\mathbb{E}\left[\left\|\delta_t\right\|^2\right] - 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,f^{'}\left(\mathbf{w}_t\right)\right\rangle + 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,\mathbb{E}\left[Z_t l^{'}\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right]\right\rangle \\ &\le \mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2\right] + \eta_t^2\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,\mathbb{E}\left[Z_t l^{'}\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right]\right\rangle \\ &\quad + 2\eta_t\left(f\left(\mathbf{w}^*\right)-f\left(\mathbf{w}_t\right)-\frac{\lambda}{2}\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2\right)\end{aligned}$$

Since the function $f\left(.\right)$ is λ-strongly convex and $\mathbf{w}^*$ is the optimal solution, we have

$$f\left(\mathbf{w}_t\right)-f\left(\mathbf{w}^*\right) \ge \left\langle f^{'}\left(\mathbf{w}_t\right),\mathbf{w}_t-\mathbf{w}^*\right\rangle + \frac{\lambda}{2}\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2 \ge \frac{\lambda}{2}\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2$$

It follows that

$$\mathbb{E}\left[\left\|\mathbf{w}_{t+1}-\mathbf{w}^*\right\|^2\right] \le \mathbb{E}\left[\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2\right] + \eta_t^2\mathbb{E}\left[\left\|\delta_t\right\|^2\right] + 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*,\mathbb{E}\left[Z_t l^{'}\left(\mathbf{w}_{t'};x_{t'},y_{t'}\right)\right]\right\rangle - 2\eta_t\lambda\left\|\mathbf{w}_t-\mathbf{w}^*\right\|^2$$

Taking expectation the above inequality, we achieve

$$\begin{aligned}
\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right] &\leq \mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]+\eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right]+2\eta_t\mathbb{E}\left[\left\langle \mathbf{w}_t-\mathbf{w}^*, Z_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\rangle\right]-2\eta_t\lambda\mathbb{E}\|\mathbf{w}_t-\mathbf{w}^*\|^2 \\
&= \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]+\eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right]+2\eta_t\mathbb{E}\left[\left\langle \mathbf{w}_t-\mathbf{w}^*, Z_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\rangle\right] \\
&\leq \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]+\eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right]+2\eta_t\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2}\mathbb{E}\left[\left\|l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\|^2\right]^{1/2}\mathbb{E}\left[Z_t^2\right]^{1/2} \\
&\leq \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]+\eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right]+2\eta_t\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2}\mathbb{E}\left[\left\|l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\|^2\right]^{1/2}\mathbb{P}\left(Z_t=1\right)^{1/2} \\
&\leq \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]+\frac{Q}{\lambda^2 t^2}+\frac{\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2}L^{1/2}}{\lambda t} \\
&\leq \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]+\frac{Q}{\lambda^2}+\frac{\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2}L^{1/2}}{\lambda t}
\end{aligned}$$

Choosing $W=\frac{\lambda L^{1/2}+\sqrt{\lambda^2 L+8\lambda^2 Q}}{4\lambda^2}$, we destine if $\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]\leq W$ then $\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right]\leq W$.

Here we note that we have bounded $\mathbb{E}\left[\|\delta_t\|^2\right]\leq 2\left(\mathbb{E}\left[\|g_t\|^2\right]+\mathbb{E}\left[\left\|l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_t'\right)\right\|^2\right]\right)=2\left(G+L\right)=Q$ and $\mathbb{E}\left[Z_t^2\right]=P\left(Z_t=1\right)=p_t\leq 1$. □

Theorem 5. *If* $\mathbf{w}^*=\underset{\mathbf{w}}{argmin}\left(\frac{\lambda}{2}\|\mathbf{w}\|^2+\frac{1}{N}\sum_{i=1}^N\left(y_i-\mathbf{w}^\intercal\Phi\left(x_i\right)\right)^2\right)$ *then* $\|\mathbf{w}^*\|\leq y_{max}\lambda^{-1/2}$.

Proof. Let us consider the equivalent constrains optimization problem

$$\min_{\mathbf{w},\boldsymbol{\xi}}\left(\frac{\lambda}{2}\|\mathbf{w}\|^2+\frac{1}{N}\sum_{i=1}^N\xi_i^2\right)$$
$$\text{s.t.:}\ \xi_i=y_i-\mathbf{w}^\intercal\Phi\left(x_i\right),\ \forall i$$

The Lagrange function is of the following form

$$\mathcal{L}\left(\mathbf{w},\boldsymbol{\xi},\boldsymbol{\alpha}\right)=\frac{\lambda}{2}\left\|\mathbf{w}^2\right\|+\frac{1}{N}\sum_{i=1}^N\xi_i^2+\sum_{i=1}^N\alpha_i\left(y_i-\mathbf{w}^\intercal\Phi\left(x_i\right)-\xi_i\right)$$

Setting the derivatives to 0, we gain

$$\nabla_{\mathbf{w}}\mathcal{L}=\lambda\mathbf{w}-\sum_{i=1}^N\alpha_i\Phi\left(x_i\right)=0\rightarrow\mathbf{w}=\lambda^{-1}\sum_{i=1}^N\alpha_i\Phi\left(x_i\right)$$

$$\nabla_{\xi_i}\mathcal{L}=\frac{2}{N}\xi_i-\alpha_i=0\rightarrow\xi_i=\frac{N\alpha_i}{2}$$

Substituting the above to the Lagrange function, we gain the dual form

$$\begin{aligned}
\mathcal{W}\left(\boldsymbol{\alpha}\right) &= -\frac{\lambda}{2}\|\mathbf{w}\|^2+\sum_{i=1}^N y_i\alpha_i-\frac{N}{4}\sum_{i=1}^N\alpha_i^2 \\
&= -\frac{1}{2\lambda}\left\|\sum_{i=1}\alpha_i\Phi\left(x_i\right)\right\|^2+\sum_{i=1}^N y_i\alpha_i-\frac{N}{4}\sum_{i=1}^N\alpha_i^2
\end{aligned}$$

Let us denote $\left(\mathbf{w}^*,\boldsymbol{\xi}^*\right)$ and $\boldsymbol{\alpha}^*$ be the primal and dual solutions, respectively. Since the strong duality holds, we have

$$\frac{\lambda}{2}\|\mathbf{w}^*\|^2+\frac{1}{N}\sum_{i=1}^N\xi_i^{*2}=-\frac{\lambda}{2}\|\mathbf{w}^*\|^2+\sum_{i=1}^N y_i\alpha_i^*-\frac{N}{4}\sum_{i=1}^N\alpha_i^{*2}$$

$$\lambda \|\mathbf{w}^*\|^2 = \sum_{i=1}^{N} y_i \alpha_i^* - \frac{N}{4} \sum_{i=1}^{N} \alpha_i^{*2} - \frac{1}{N} \sum_{i=1}^{N} \xi_i^{*2}$$
$$\leq \sum_{i=1}^{N} \left(y_i \alpha_i^* - \frac{N}{4} \alpha_i^{*2} \right) \leq \sum_{i=1}^{N} \frac{y_i^2}{N} \leq y_{\max}^2$$

We note that we have used $g(\alpha_i^*) = y_i \alpha_i^* - \frac{N}{4}\alpha_i^{*2} \leq g\left(\frac{2y_i}{N}\right) = \frac{y_i^2}{N}$. Hence, we gain the conclusion. □

Lemma 6. *Assume that L2 loss is using, the following statement holds*

$$\|\mathbf{w}_{T+1}\| \leq \lambda^{-1} \left(y_{max} + \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{w}_t\| \right)$$

where $y_{max} = \max_{y \in \mathcal{Y}} |y|$.

Proof. We have the following

$$\mathbf{w}_{t+1} = \prod_{S} \left(\frac{t-1}{t} \mathbf{w}_t - \eta_t \alpha_t \Phi(x_t) \right)$$

It follows that

$$\|\mathbf{w}_{t+1}\| \leq \frac{t-1}{t} \|\mathbf{w}_t\| + \frac{1}{\lambda t} |\alpha_t| \quad \text{since } \|\Phi(x_t)\| = 1$$

It happens that $l'(\mathbf{w}_t; x_t, y_t) = \alpha_t \Phi(x_t)$. Hence, we gain

$$|\alpha_t| = \left| y_t - \mathbf{w}_t^{\mathsf{T}} \Phi(x_t) \right| \leq y_{\max} + \|\mathbf{w}_t\| \|\Phi(x_t)\| \leq y_{\max} + \|\mathbf{w}_t\|$$

It implies that

$$t \|\mathbf{w}_{t+1}\| \leq (t-1) \|\mathbf{w}_t\| + \lambda^{-1} (y_{\max} + \|\mathbf{w}_t\|)$$

Taking sum when $t = 1, 2, \ldots, T$, we achieve

$$T \|\mathbf{w}_{T+1}\| \leq \lambda^{-1} \left(T y_{\max} + \sum_{t=1}^{T} \|\mathbf{w}_t\| \right)$$

$$\|\mathbf{w}_{T+1}\| \leq \lambda^{-1} \left(y_{\max} + \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{w}_t\| \right) \tag{1}$$

□

Theorem 7. *If* $\lambda > 1$ *then* $\|\mathbf{w}_{T+1}\| \leq \frac{y_{max}}{\lambda - 1}\left(1 - \frac{1}{\lambda^T}\right) < \frac{y_{max}}{\lambda - 1}$ *for all* T.

Proof. First we consider the sequence $\{s_T\}_T$ which is identified as $s_{T+1} = \lambda^{-1}(y_{\max} + s_T)$ and $s_1 = 0$. It is easy to find the formula of this sequence as

$$s_{T+1} - \frac{y_{\max}}{\lambda - 1} = \lambda^{-1} \left(s_T - \frac{y_{\max}}{\lambda - 1} \right) = \ldots = \lambda^{-T} \left(s_1 - \frac{y_{\max}}{\lambda - 1} \right) = \frac{\lambda^{-T} y_{\max}}{\lambda - 1}$$

$$s_{T+1} = \frac{y_{\max}}{\lambda - 1} \left(1 - \frac{1}{\lambda^T} \right)$$

We prove by induction by T that $\|\mathbf{w}_T\| \leq s_T$ for all T. It is obvious that $\|\mathbf{w}_1\| = s_1 = 0$. Assume that $\|\mathbf{w}_t\| \leq s_t$ for $t \leq T$, we verify it for $T+1$. Indeed, we have

$$\|\mathbf{w}_{T+1}\| \leq \lambda^{-1} \left(y_{\max} + \frac{1}{T} \sum_{t=1}^{T} \|\mathbf{w}_t\| \right) \leq \lambda^{-1} \left(y_{\max} + \frac{1}{T} \sum_{t=1}^{T} s_t \right)$$
$$\leq \lambda^{-1} (y_{\max} + s_T) = s_{T+1}$$

□

Theorem 8. *Let us consider running of Algorithm 2 where (x_t, y_t) is sampled from the training set $\mathcal{D}$ or the join distribution $\mathbb{P}_{X,Y}$. Let define the gradient error as $M_t = \frac{\Delta_t}{\eta_t} = -l^{'}(\mathbf{w}_{t'}; x_{t'}, y_{t'})$. We have the following*

$$\mathbb{E}\left[f\left(\overline{\mathbf{w}}_T\right) - f\left(\mathbf{w}^*\right)\right] \le \frac{Q\left(\log T + 1\right)}{2\lambda T} + \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2} \mathbb{P}\left(Z_t = 1\right)^{1/2}$$
$$\le \frac{Q\left(\log T + 1\right)}{2\lambda T} + \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2}$$

Proof. Let us define $\delta_t = g_t + Z_t M_t$. We have $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \delta_t$.

$$\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2 = \left\|\mathbf{w}_t - \eta_t \delta_t - \mathbf{w}^*\right\|^2 = \left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2 + \eta_t^2 \left\|\delta_t\right\|^2 - 2\eta_t \left\langle \mathbf{w}_t - \mathbf{w}^*, \delta_t \right\rangle$$

$$\left\langle \mathbf{w}_t - \mathbf{w}^*, g_t \right\rangle = \frac{\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2 - \left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2}{2\eta_t} + \frac{\eta_t \left\|\delta_t\right\|^2}{2} - \left\langle \mathbf{w}_t - \mathbf{w}^*, Z_t M_t \right\rangle$$

Taking the conditional expectation w.r.t $\mathbf{w}_t$, we achieve

$$\left\langle \mathbf{w}_t - \mathbf{w}^*, \mathbb{E}\left[g_t\right] \right\rangle = \frac{\mathbb{E}\left[\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2\right]}{2\eta_t} + \frac{\eta_t \mathbb{E}\left[\left\|\delta_t\right\|^2\right]}{2} - \left\langle \mathbf{w}_t - \mathbf{w}^*, \mathbb{E}\left[Z_t M_t\right] \right\rangle$$

$$\left\langle \mathbf{w}_t - \mathbf{w}^*, f^{'}\left(\mathbf{w}_t\right) \right\rangle = \frac{\mathbb{E}\left[\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2\right]}{2\eta_t} + \frac{\eta_t \mathbb{E}\left[\left\|\delta_t\right\|^2\right]}{2} - \left\langle \mathbf{w}_t - \mathbf{w}^*, \mathbb{E}\left[Z_t M_t\right] \right\rangle$$

$$f\left(\mathbf{w}_t\right) - f\left(\mathbf{w}^*\right) + \frac{\lambda}{2}\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2 \le \frac{\mathbb{E}\left[\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\right] - \mathbb{E}\left[\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2\right]}{2\eta_t}$$
$$+ \frac{\eta_t \mathbb{E}\left[\left\|\delta_t\right\|^2\right]}{2} - \left\langle \mathbf{w}_t - \mathbf{w}^*, \mathbb{E}\left[Z_t M_t\right] \right\rangle$$

Taking expectation, we come to the following

$$\mathbb{E}\left[f\left(\mathbf{w}_t\right) - f\left(\mathbf{w}^*\right)\right] \le \frac{\lambda}{2}\left(t-1\right)\mathbb{E}\left[\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\right] - \frac{\lambda}{2} t \mathbb{E}\left[\left\|\mathbf{w}_{t+1} - \mathbf{w}^*\right\|^2\right] + \frac{Q}{2\lambda t}$$
$$+ E\left[\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\right]^{1/2} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2} \mathbb{E}\left[Z_t^2\right]^{1/2}$$

Summing when $t = 1, 2, \dots, T$, we gain

$$\mathbb{E}\left[\frac{\sum_{t=1}^{T} f\left(\mathbf{w}_t\right)}{T} - f\left(\mathbf{w}^*\right)\right] \le \frac{Q}{2\lambda T}\sum_{t=1}^{T}\frac{1}{t} + \frac{1}{T}\sum_{t=1}^{T} E\left[\left\|\mathbf{w}_t - \mathbf{w}^*\right\|^2\right]^{1/2} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2} \mathbb{E}\left[Z_t^2\right]^{1/2}$$
$$\le \frac{Q\left(\log T + 1\right)}{2\lambda T} + \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2} \mathbb{P}\left(Z_t = 1\right)^{1/2}$$
$$\le \frac{Q\left(\log T + 1\right)}{2\lambda T} + \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2} \quad (2)$$

Let $\overline{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$, we reach

$$\mathbb{E}\left[f\left(\overline{\mathbf{w}}_T\right) - f\left(\mathbf{w}^*\right)\right] \le \frac{Q\left(\log T + 1\right)}{2\lambda T} + \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2} \mathbb{P}\left(Z_t = 1\right)^{1/2}$$
$$\le \frac{Q\left(\log T + 1\right)}{2\lambda T} + \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\left\|M_t\right\|^2\right]^{1/2}$$

□

Theorem 9. *We denote the gap*

$$d_T = \frac{1}{T} W^{1/2} \sum_{t=1}^{T} \mathbb{E}\left[\|M_t\|^2\right]^{1/2} \mathbb{P}\left(Z_t = 1\right)^{1/2}$$

Let r be an integer picked uniformly at random from $\{1, 2, \ldots, T\}$. Then, with probability of at least $1 - \delta$ we have

$$f\left(\mathbf{w}_r\right) \leq f\left(\mathbf{w}^*\right) + d_T + \frac{Q\left(\log T + 1\right)}{2\lambda T \delta}$$

Proof. Let us denote $X = f\left(\mathbf{w}_r\right) - f\left(\mathbf{w}^*\right) \geq 0$ and $Y = \frac{\sum_{t=1}^{T} f(\mathbf{w}_t)}{T} - f\left(\mathbf{w}^*\right)$. Then, we have

$$\mathbb{E}_r\left[X\right] = \mathbb{E}_r\left[f\left(\mathbf{w}_r\right) - f\left(\mathbf{w}^*\right)\right] = \frac{\sum_{t=1}^{T} f\left(\mathbf{w}_t\right)}{T} - f\left(\mathbf{w}^*\right) = Y$$

Therefore, we gain

$$\mathbb{E}\left[X\right] = \mathbb{E}_{(x_t, y_t)_1^T}\left[\mathbb{E}_r\left[X\right]\right] = \mathbb{E}\left[Y\right] \leq \frac{Q\left(\log T + 1\right)}{2\lambda T} + d_T$$

or equivalently

$$\mathbb{E}\left[X - d_T\right] = \mathbb{E}\left[Y - d_T\right] \leq \frac{Q\left(\log T + 1\right)}{2\lambda T}$$

where $(x_t, y_t)_1^T$ specifies the sequence of incoming instances $\{(x_1, y_1), \ldots, (x_T, y_T)\}$ and we refer to Eq. (2) for last inequality.

According to Markov inequality, we have

$$\mathbb{P}\left(X - d_T \geq \varepsilon\right) \leq \frac{\mathbb{E}\left[X - d_T\right]}{\varepsilon} \leq \frac{Q\left(\log T + 1\right)}{2\lambda T \varepsilon}$$

$$\mathbb{P}\left(X - d_T < \varepsilon\right) \geq 1 - \frac{Q\left(\log T + 1\right)}{2\lambda T \varepsilon}$$

Choosing $\varepsilon = \frac{Q(\log T + 1)}{2\lambda T \delta}$, we obtain the conclusion. ☐

Corollary 10. *If $\mathbb{E}\left[Z_t^2\right] = \mathbb{P}\left(Z_t = 1\right) = p_t \sim O\left(\frac{1}{t}\right)$ then $\mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2\right] \sim O\left(\frac{1}{t}\right)$.*

Proof. Let us define $\delta_t = g_t - Z_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)$. We have the following

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \delta_t$$

$$\begin{aligned} \|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2 &= \|\mathbf{w}_t - \eta_t \delta_t - \mathbf{w}^*\|^2 = \|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_t^2 \|\delta_t\|^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}^*, \delta_t \rangle \\ &= \|\mathbf{w}_t - \mathbf{w}^*\|^2 + \eta_t^2 \|\delta_t\|^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}^*, g_t \rangle + 2\eta_t \left\langle \mathbf{w}_t - \mathbf{w}^*, Z_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right) \right\rangle \end{aligned}$$

Taking conditional expectation w.r.t $\mathbf{w}_t^1$, x_1^{t-1} and note that $t' < t$, we gain

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|^2\right] &= \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2\right] + \eta_t^2 \mathbb{E}\left[\|\delta_t\|^2\right] - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}^*, \mathbb{E}\left[g_t\right] \rangle + 2\eta_t \left\langle \mathbf{w}_t - \mathbf{w}^*, l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right) \mathbb{E}\left[Z_t\right] \right\rangle \\ &= \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2\right] + \eta_t^2 \mathbb{E}\left[\|\delta_t\|^2\right] - 2\eta_t \left\langle \mathbf{w}_t - \mathbf{w}^*, f^{'}\left(\mathbf{w}_t\right) \right\rangle + 2\eta_t \left\langle \mathbf{w}_t - \mathbf{w}^*, p_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right) \right\rangle \\ &\leq \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2\right] + \eta_t^2 \mathbb{E}\left[\|\delta_t\|^2\right] + 2\eta_t \left\langle \mathbf{w}_t - \mathbf{w}^*, p_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right) \right\rangle \\ &\quad + 2\eta_t \left(f\left(\mathbf{w}^*\right) - f\left(\mathbf{w}_t\right) - \frac{\lambda}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 \right) \end{aligned}$$

Since the function $f\left(.\right)$ is λ-strongly convex and $\mathbf{w}^*$ is the optimal solution, we have

$$f\left(\mathbf{w}_t\right) - f\left(\mathbf{w}^*\right) \geq \left\langle f^{'}\left(\mathbf{w}_t\right), \mathbf{w}_t - \mathbf{w}^* \right\rangle + \frac{\lambda}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2 \geq \frac{\lambda}{2} \|\mathbf{w}_t - \mathbf{w}^*\|^2$$

It follows that

$$\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right] \le \mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right] + 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*, p_t l\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\rangle - 2\eta_t\lambda\|\mathbf{w}_t-\mathbf{w}^*\|^2$$

Taking expectation the above inequality, we achieve

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right] &\le \mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right] + 2\eta_t\mathbb{E}\left[\left\langle \mathbf{w}_t-\mathbf{w}^*, p_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\rangle\right] - 2\eta_t\lambda\mathbb{E}\left\|[\mathbf{w}_t-\mathbf{w}^*\right\|^2 \\ &= \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right] + 2\eta_t\mathbb{E}\left[\left\langle \mathbf{w}_t-\mathbf{w}^*, p_t l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\rangle\right] \\ &\le \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right] + 2\eta_t\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2}\mathbb{E}\left[p_t^2\left\|l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\|^2\right]^{1/2}\end{aligned}$$

Since $p_t \sim O\left(\frac{1}{t}\right)$, we have $p_t < \frac{C}{t}$ for some $C > 0$. Therefore, the above inequality becomes

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right] &\le \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] + \eta_t^2\mathbb{E}\left[\|\delta_t\|^2\right] + 2\eta_t\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2}\frac{C}{t}\mathbb{E}\left[\left\|l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)\right\|\right]^{1/2} \\ &\le \frac{t-2}{t}\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] + \frac{Q}{\lambda^2 t^2} + \frac{\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right]^{1/2} CL^{1/2}}{\lambda t^2}\end{aligned}$$

By choosing $W_t = \frac{Q^2\lambda^{-2}+M^{1/2}CL^{1/2}}{t}$, we gain if $\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] \le W_t$, then $\mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right] \le W_{t+1}$. □

Theorem 11. *Let us consider running of Algorithm 3 where* (x_t, y_t) *is sampled from the training set* $\mathcal{D}$ *or the join distribution* $\mathbb{P}_{X,Y}$. *Let define the gradient error as* $M_t = \frac{\Delta_t}{\eta_t} = -l^{'}\left(\mathbf{w}_{t'}; x_{t'}, y_{t'}\right)$. *We have the following*

$$\mathbb{E}\left[f\left(\overline{\mathbf{w}}_T^{\gamma}\right) - f\left(\mathbf{w}^*\right)\right] \le \frac{D\lambda^2 + Q\log\left(1/\left(1-\gamma\right)\right)}{2\gamma T} + \frac{\beta D^{1/2}}{\gamma T}\sum_{t=(1-\gamma)T+1}^{T}\frac{\mathbb{E}\left[\|M_t\|^2\right]^{1/2}}{t^{3/2}} \tag{3}$$

Proof. Let us define $\delta_t = g_t + Z_t M_t$. We have the following

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t\delta_t$$

$$\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2 = \|\mathbf{w}_t - \eta_t\delta_t - \mathbf{w}^*\|^2 = \|\mathbf{w}_t-\mathbf{w}^*\|^2 + \eta_t^2\|\delta_t\|^2 - 2\eta_t\left\langle \mathbf{w}_t-\mathbf{w}^*, \delta_t\right\rangle$$

$$\left\langle \mathbf{w}_t-\mathbf{w}^*, g_t\right\rangle = \frac{\|\mathbf{w}_t-\mathbf{w}^*\|^2 - \|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2}{2\eta_t} + \frac{\eta_t\|\delta_t\|^2}{2} - \left\langle \mathbf{w}_t-\mathbf{w}^*, Z_t M_t\right\rangle$$

Taking the conditional expectation w.r.t $\mathbf{w}_t^1$, we achieve

$$\left\langle \mathbf{w}_t-\mathbf{w}^*, \mathbb{E}\left[g_t\right]\right\rangle = \frac{\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right]}{2\eta_t} + \frac{\eta_t\mathbb{E}\left[\|\delta_t\|^2\right]}{2} - \left\langle \mathbf{w}_t-\mathbf{w}^*, \mathbb{E}\left[Z_t M_t\right]\right\rangle$$

$$\left\langle \mathbf{w}_t-\mathbf{w}^*, f^{'}\left(\mathbf{w}_t\right)\right\rangle = \frac{\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right]}{2\eta_t} + \frac{\eta_t\mathbb{E}\left[\|\delta_t\|^2\right]}{2} - \left\langle \mathbf{w}_t-\mathbf{w}^*, \mathbb{E}\left[Z_t M_t\right]\right\rangle$$

$$f\left(\mathbf{w}_t\right) - f\left(\mathbf{w}^*\right) \le \frac{\mathbb{E}\left[\|\mathbf{w}_t-\mathbf{w}^*\|^2\right] - \mathbb{E}\left[\|\mathbf{w}_{t+1}-\mathbf{w}^*\|^2\right]}{2\eta_t} + \frac{\eta_t\mathbb{E}\left[\|\delta_t\|^2\right]}{2} - \left\langle \mathbf{w}_t-\mathbf{w}^*, \mathbb{E}\left[Z_t M_t\right]\right\rangle$$

Taking expectation and summing when $t = (1-\gamma)T+1,\dots,T$, let $\overline{\mathbf{w}}_T^{\gamma} = \frac{1}{\gamma T}\sum_{t=(1-\gamma)T+1}^{T}\mathbf{w}_t$ and note that $p_t \le P\left(S_t = 1\right) \le \frac{\beta}{t}$, we reach the following

$$\gamma T\mathbb{E}\left[\frac{\sum_{t=(1-\gamma)T+1}^{T} f(\mathbf{w}_t)}{\gamma T} - f(\mathbf{w}^*)\right] \le \frac{\mathbb{E}\left[\left\|\mathbf{w}_{(1-\gamma)T+1} - \mathbf{w}^*\right\|^2\right]}{2\eta_{(1-\gamma)T+1}} + \sum_{t=(1-\gamma)T+2}^{T} \mathbb{E}\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2\right]\left(\frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}}\right) \tag{4}$$

$$+ \sum_{t=(1-\gamma)T+1}^{T}\left(\frac{\eta_t \mathbb{E}\left[\|\delta_t\|^2\right]}{2} + E\left[\|\mathbf{w}_t - \mathbf{w}^*\|^2\right]^{1/2}\mathbb{E}\left[\|M_t\|^2\right]^{1/2} p_t\right)$$

$$\le \frac{W_T\lambda((1-\gamma)T+1)}{2} + \frac{W_T\lambda(\gamma T - 1)}{2} + \frac{Q}{2\lambda}\sum_{t=(1-\gamma)T+1}^{T}\frac{1}{t} + \sum_{t=(1-\gamma)T+1}^{T} W_t^{1/2}\mathbb{E}\left[\|M_t\|^2\right]^{1/2}\frac{\beta}{t}$$

$$\le \frac{W_T\lambda T}{2} + + \frac{Q}{2\lambda}\sum_{t=(1-\gamma)T+1}^{T}\frac{1}{t} + \beta D^{1/2}\sum_{t=(1-\gamma)T+1}^{T}\frac{\mathbb{E}\left[\|M_t\|^2\right]^{1/2}}{t^{3/2}}$$

$$\le \frac{D\lambda}{2} + \frac{Q\log(1/(1-\gamma))}{2\lambda} + \beta D^{1/2}\sum_{t=(1-\gamma)T+1}^{T}\frac{\mathbb{E}\left[\|M_t\|^2\right]^{1/2}}{t^{3/2}}$$

$$\gamma T\mathbb{E}\left[f(\overline{\mathbf{w}}_T^{\gamma}) - f(\mathbf{w}^*)\right] \le \frac{D\lambda}{2} + \frac{Q\log(1/(1-\gamma))}{2\lambda} + \beta D^{1/2}\sum_{t=(1-\gamma)T+1}^{T}\frac{\mathbb{E}\left[\|M_t\|^2\right]^{1/2}}{t^{3/2}}$$

To derive the last inequality, we use the facts $\sum_{t=(1-\gamma)T+1}^{T}\frac{1}{t} \le \log(1/(1-\gamma))$ and $W_t \le \frac{D}{t}$ for all t. Finally, we achieve

$$\mathbb{E}\left[f(\overline{\mathbf{w}}_T^{\gamma}) - f(\mathbf{w}^*)\right] \le \frac{D\lambda^2 + Q\log(1/(1-\gamma))}{2\gamma T} + \frac{\beta D^{1/2}}{\gamma T}\sum_{t=(1-\gamma)T+1}^{T}\frac{\mathbb{E}\left[\|M_t\|^2\right]^{1/2}}{t^{3/2}}$$

□

Theorem 12. *Let us consider running of Algorithm 3 where (x_t, y_t) is sampled from the training set $\mathcal{D}$ or the join distribution $\mathbb{P}_{X,Y}$. We have the following*

$$\mathbb{E}\left[f(\overline{\mathbf{w}}_T^{\gamma}) - f(\mathbf{w}^*)\right] \le \frac{D\lambda^2 + Q\log(1/(1-\gamma)) + 2\beta L D^{1/2}\log(1/(1-\gamma))}{2\gamma T}$$

Proof. To gain the conclusion, we use inequality in Eq. (3) and note that $\mathbb{E}\left[\|M_t\|^2\right]^{1/2} = \mathbb{E}\left[\left\|l'(\mathbf{w}_{t'}; x_{t'}, y_{t'})\right\|^2\right]^{1/2} \le L$. □

Theorem 13. *Let r be an integer randomly picked from $\{(1-\gamma)T+1, \ldots, T\}$. Then, with probability at least $1-\delta$, we have*

$$f(\mathbf{w}_r) \le f(\mathbf{w}^*) + \frac{R}{2\gamma\delta T}$$

where we have defined $R = D\lambda^2 + Q\log(1/(1-\gamma)) + 2\beta L D^{1/2}\log(1/(1-\gamma))$.

Proof. Let us denote $X = f(\mathbf{w}_r) - f(\mathbf{w}^*) \ge 0$ and $Y = \frac{\sum_{t=(1-\gamma)T+1}^{T} f(\mathbf{w}_t)}{\gamma T} - f(\mathbf{w}^*)$. Then, we have

$$\mathbb{E}_r[X] = \mathbb{E}_r\left[f(\mathbf{w}_r) - f(\mathbf{w}^*)\right] = \frac{\sum_{t=(1-\gamma)T+1}^{T} f(\mathbf{w}_t)}{\gamma T} - f(\mathbf{w}^*) = Y$$

Therefore, we gain

$$\mathbb{E}[X] = \mathbb{E}_{(x_t,y_t)_1^T}[\mathbb{E}_r[X]] = \mathbb{E}[Y] \leq \frac{R}{2\gamma T} \tag{5}$$

Note that to achieve the last inequality in Eq. (5), we refer to Eq. (4).
According to Markov inequality, we have

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon} \leq \frac{R}{2\gamma T}$$

$$\mathbb{P}(X < \varepsilon) \geq 1 - \frac{R}{2\gamma T}$$

Choosing $\varepsilon = \frac{R}{2\gamma\delta T}$, we gain the conclusion. □

4 Exact Projection

We present in detail how to incrementally maintain the inverse matrix K_t^{-1}. We consider two cases

- $|\mathcal{I}_t| \leq B$
  We compute as follows:
  Compute $d = K_{t-1}^{-1} k_t$
  Set $\|\delta_t\|^2 = K(x_t, x_t) - k_t^\mathsf{T} d$
  Update

$$K_t^{-1} = \begin{bmatrix} & K_{t-1}^{-1} & & 0 \\ & & & \dots \\ & & & 0 \\ 0 & \dots & 0 & 0 \end{bmatrix} + \frac{1}{\|\delta_t\|^2} \begin{bmatrix} d \\ -1 \end{bmatrix} \begin{bmatrix} d^\mathsf{T} & -1 \end{bmatrix}$$

  The computational cost to maintain K_t^{-1} when t varies from 1 to B is $\sum_{t=1}^{B} \mathrm{O}(t^2) = \mathrm{O}(B^3)$.

- $|\mathcal{I}_t| = B + 1$
  To update K_t^{-1} from K_{t-1}^{-1} we observe that these two matrices K_{t-1} and K_t are distinct in one row and one column. Concretely, to transform K_{t-1} to K_t, we can substitute the column $\boldsymbol{k}_p$ by $\boldsymbol{k}_t$ and do the same for the corresponding row. Therefore, we can formulate $K_t = K_{t-1} + L$ where L is a sparse matrix of all zeros except for one column and row, which can be computed as $L_p = \boldsymbol{k}_t - \boldsymbol{k}_p$. It is apparent that $rank(L) = 2$. To update K_t^{-1} from K_{t-1}, we rely on Thm. 14 (cf. [1]).

  We assume that the i-th collumn and row in $B \times B$ matrice K_{t-1} and K_t is mapped to the element $x_{\pi(i)}$ in $\{x_1, x_2, \dots, x_t\}$. We further assume the removal element x_p locates at m-th collumn in matrix K_{t-1}. To gain K_t from K_{t-1}, we replace x_p by x_t and hence $\pi^{-1}(t) = \pi^{-1}(p) = m$. It is evident that $K_t = K_{t-1} + L$ where L is a matrix of all zeros except for $m-$th column and row, which is computed as $L_m(i) = K(x_t, x_{\pi(i)}) - K(x_p, x_{\pi(i)})$ for $i = 1, \dots, B$. It is apparent that $rank(L) = 2$ and it can be decomposed as $L = L_1 + L_2$ where L_1, L_2 are matrices of all zeros except for m-th column and m-th row respectively and hence $rank(L_1) = rank(L_2) = 1$.
  To directly apply Thm. 14, we denote $C_1 = A = K_{t-1}$, $B_1 = L_1$, and $B_2 = L_2$. We first compute C_2^{-1} by

$$C_2^{-1} = C_1^{-1} - g_1 C_1^{-1} B_1 C_1^{-1} \tag{6}$$

  It is obvious the computational cost to compute C_2^{-1} as in Eq. (6) is $\mathrm{O}(B^2)$.
  We then compute $K_t^{-1} = (A+B)^{-1} = (A + B_1 + B_2)^{-1}$ as

$$K_t^{-1} = (A+B)^{-1} = C_2^{-1} - g_2 C_2^{-1} B_2 C_2^{-1} \tag{7}$$

  The computional cost of Eq. (7) is again $\mathrm{O}(B^2)$.

Theorem 14. *Let A and $A+B$ be nonsingular matrices, and let B have rank $r > 0$. Let $B = B_1 + \dots + B_r$, where each B_i has rank 1, and each $C_{k+1} = A + B_1 + \dots + B_k$ is nonsingular. Setting $C_1 = A$, then $C_{k+1}^{-1} = C_k^{-1} - g_k C_k^{-1} B_k C_k^{-1}$ where $g_k = \frac{1}{1+trace(C_k^{-1}B_k)}$. In particular, $(A+B)^{-1} = C_r^{-1} - g_r C_r^{-1} B_r C_r^{-1}$.*

References

[1] K. S. Miller. On the Inverse of the Sum of Matrices. *Mathematics Magazine*, 54(2):67–72, 1981.