Supplementary Material for Nonparametric Budgeted Stochastic Gradient Descent

1 Notion

We introduce some notions used in this supplementary material. For regression task, we define $y_{\text{max}} = \max_y |y|$. We further denote the set $S$ as

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

where $\mathcal{B}(0, y_{\text{max}} \lambda^{-1/2}) = \{ w \in \mathbb{R}^D : \|w\| \leq y_{\text{max}} \lambda^{-1/2} \}$ 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, $\|l'(w; x, y)\| \leq A \|w\|^{1/2} + B$ 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**

  $$l(w; x, y) = \max \{0, 1 - yw^T\Phi(x)\}$$

  $$l'(w; x, y) = -I\{yw^T\Phi(x) \leq 1\} y\Phi(x)$$

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

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

- **L2 loss**

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

  $$l(w; x, y) = \frac{1}{2} (y - w^T\Phi(x))^2$$

  $$l'(w; x, y) = (w^T\Phi(x) - y) \Phi(x)$$

  $$\|l'(w_t; x, y)\| = \|w_t^T\Phi(x) + y\| \|\Phi(x)\| \leq |w_t^T\Phi(x)| + y_{\max}$$

  $$\leq \|\Phi(x)\| \|w_t\| + y_{\max} \leq A \|w_t\|^{1/2} + B$$

  where $B = y_{\max}$ and $A = \begin{cases} \frac{\lambda^{-1/2}}{2} & \text{if } \lambda \leq 1 \\ \frac{\lambda^{-1/2}}{2} & \text{otherwise} \end{cases}$. Here we note that we make use of the fact that $\|w_t\| \leq y_{\max} (\lambda - 1)^{-1}$ if $\lambda > 1$ (cf. Thm. 7) and $\|w_t\| \leq y_{\max} \lambda^{-1/2}$ otherwise (cf. Line 13 in Alg. 2 and Line 16 in Alg. 3).
• **L1 loss**

\[
l(w; x, y) = |y - w^T \Phi(x)|
\]

\[
l'(w; x, y) = \text{sign}(w^T \Phi(x) - y) \Phi(x)
\]

Therefore, by choosing \( A = 0, B = 1 \) we have

\[
\|l'(w; x, y)\| = \|\Phi(x)\| \leq A \|w\|^{1/2} + B
\]

• **Logistic loss**

\[
l(w; x, y) = \log(1 + \exp(-y w^T \Phi(x)))
\]

\[
l'(w; x, y) = \frac{-y \exp(-y w^T \Phi(x)) \Phi(x)}{\exp(-y w^T \Phi(x)) + 1}
\]

Therefore, by choosing \( A = 0, B = 1 \) we have

\[
\|l'(w; x, y)\| < \|\Phi(x)\| \leq A \|w\|^{1/2} + B
\]

• **\( \varepsilon \)-insensitive loss**

\[
l(w; x, y) = \max\{0, |y - w^T \Phi(x)| - \varepsilon\}
\]

\[
l'(w; x, y) = \mathbb{I}_{(|y - w^T \Phi(x)| > \varepsilon)} \text{sign}(w^T \Phi(x) - y)x
\]

Therefore, by choosing \( A = 0, B = 1 \) we have

\[
\|l'(w; x, y)\| = \|\Phi(x)\| \leq A \|w\|^{1/2} + B
\]

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 \( t \),

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

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

It is obvious for \( t = 1 \) from \( \mathbb{E}\left[\|w_1\|^2\right]^{1/2} = 0 \).

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

\[
\sqrt{\mathbb{E}\left[\|w_{t+1}\|^2\right]} \leq \frac{t-1}{t} \sqrt{\mathbb{E}\left[\|w_t\|^2\right]} + \frac{1}{\lambda t} \sqrt{\mathbb{E}\left[\|l'(w_t; x_t, y_t)\|^2\right]} + \frac{1}{\lambda t} \sqrt{\mathbb{E}\left[Z_t^2 \|l'(w_t; x_t, y_t)\|^2\right]}
\]

\[
\leq \frac{t-1}{t} \sqrt{\mathbb{E}\left[\|w_t\|^2\right]} + \frac{1}{\lambda t} \sqrt{\mathbb{E}\left[\|l'(w_t; x_t, y_t)\|^2\right]} + \frac{1}{\lambda t} \sqrt{\mathbb{E}\left[l'(w_t; x_t, y_t)\|^2\right]}
\]

\[
\leq \frac{t-1}{t} \sqrt{\mathbb{E}\left[\|w_t\|^2\right]} + \frac{1}{\lambda t} \left(A \sqrt{\mathbb{E}\|w_t\|} + B + A \sqrt{\mathbb{E}\|w_{t'}\|} + B\right)
\]

\[
\leq \frac{t-1}{t} P + \frac{2}{\lambda t} \left(A \sqrt{P} + B\right) = P
\]

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

**Corollary 2.** The following holds for all \( t \),

\[
\mathbb{E}\left[\|l'(w_t; x_t, y_t)\|^2\right] \leq L = \left(A \sqrt{P} + B\right)^2
\]
Proof. We have the following

\[
\sqrt{E \left[ \| l'(w_t; x_t, y_t) \|^2 \right]} \leq \sqrt{E \left[ (A \| w_t \|^1/2 + B)^2 \right]} \leq A \sqrt{E \| w_t \|} + B \leq A \sqrt{P} + B
\]

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

Corollary 3. The following holds for all \( t \),

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

Proof. Again using Minkowski inequality

\[
\sqrt{E \left[ \| g_t \|^2 \right]} \leq \lambda \sqrt{E \left[ \| w_t \|^2 \right]} + \sqrt{E \left[ \| l'(w_t; x_t, y_t) \|^2 \right]} \leq \lambda P + A \sqrt{P} + B
\]

Corollary 4. The following holds for all \( t \),

\[
E \left[ \| w_t - w^* \|^2 \right] \leq W = \frac{\lambda L^{1/2} + \sqrt{\lambda^2 L + 8 \lambda^2 Q}}{4 \lambda^2}
\]

Proof. Let us define \( \delta_t = g_t - Z_t l'(w_{t'}; x_{t'}, y_{t'}) \). We have the following

\[
\| w_{t+1} - w^* \|^2 = \| w_t - \eta \delta_t - w^* \|^2 = \| w_t - w^* \|^2 + \eta^2 \| \delta_t \|^2 - 2 \eta \langle w_t - w^*, \delta_t \rangle
\]

Taking conditional expectation w.r.t \( w_1 \), we gain

\[
E \left[ \| w_{t+1} - w^* \|^2 \right] = E \left[ \| w_t - w^* \|^2 \right] + E \left[ \| \delta_t \|^2 \right] - 2 \eta \langle w_t - w^*, E \left[ g_t \right] \rangle + 2 \eta \langle w_t - w^*, E \left[ Z_t l'(w_{t'}; x_{t'}, y_{t'}) \right] \rangle
\]

Since the function \( f(.) \) is \( \lambda \)-strongly convex and \( w^* \) is the optimal solution, we have

\[
f(w_t) - f(w^*) \geq \langle f'(w_t), w_t - w^* \rangle + \frac{\lambda}{2} \| w_t - w^* \|^2 \geq \frac{\lambda}{2} \| w_t - w^* \|^2
\]

It follows that

\[
E \left[ \| w_{t+1} - w^* \|^2 \right] \leq E \left[ \| w_t - w^* \|^2 \right] + \eta^2 E \left[ \| \delta_t \|^2 \right] + 2 \eta \langle w_t - w^*, E \left[ Z_t l'(w_{t'}; x_{t'}, y_{t'}) \right] \rangle - 2 \eta \lambda \| w_t - w^* \|^2
\]
Theorem 5. If \( w^* = \arg \min_w \left( \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{i=1}^N (y_i - w^T \Phi(x_i))^2 \right) \) then \( \|w^*\| \leq y_{\max} \lambda^{-1/2}. \)

**Proof.** Let us consider the equivalent constrains optimization problem
\[
\min_{w, \xi} \left( \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{i=1}^N \xi_i^2 \right)
\text{s.t.: } \xi_i = y_i - w^T \Phi(x_i), \ \forall i
\]

The Lagrange function is of the following form
\[
L(w, \xi, \alpha) = \frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_{i=1}^N \xi_i^2 + \sum_{i=1}^N \alpha_i (y_i - w^T \Phi(x_i) - \xi_i)
\]

Setting the derivatives to 0, we gain
\[
\nabla_w L = \lambda w - \sum_{i=1}^N \alpha_i \Phi(x_i) = 0 \rightarrow w = \lambda^{-1} \sum_{i=1}^N \alpha_i \Phi(x_i)
\]
\[
\nabla_{\xi_i} L = \frac{2}{N} \xi_i - \alpha_i = 0 \rightarrow \xi_i = \frac{N \alpha_i}{2}
\]

Substituting the above to the Lagrange function, we gain the dual form
\[
W(\alpha) = -\frac{\lambda}{2} \|w\|^2 + \sum_{i=1}^N y_i \alpha_i - \frac{N}{4} \sum_{i=1}^N \alpha_i^2
\]
\[
= -\frac{1}{2\lambda} \left\| \sum_{i=1}^N \alpha_i \Phi(x_i) \right\|^2 + \sum_{i=1}^N y_i \alpha_i - \frac{N}{4} \sum_{i=1}^N \alpha_i^2
\]

Let us denote \((w^*, \xi^*)\) and \(\alpha^*\) be the primal and dual solutions, respectively. Since the strong duality holds, we have
\[
\frac{\lambda}{2} \|w^*\|^2 + \frac{1}{N} \sum_{i=1}^N \xi_i^2 = -\frac{\lambda}{2} \|w^*\|^2 + \sum_{i=1}^N y_i \alpha_i^* - \frac{N}{4} \sum_{i=1}^N \alpha_i^2
\]
\[
\lambda \| \mathbf{w}^* \|^2 = \sum_{i=1}^{N} y_i \alpha_i^* - \frac{N}{4} \sum_{i=1}^{N} \alpha_i^{*2} - \frac{1}{N} \sum_{i=1}^{N} \xi_i^2 \leq \sum_{i=1}^{N} \left( y_i \alpha_i^* - \frac{N}{4} \alpha_i^{*2} \right) \leq \sum_{i=1}^{N} \frac{y_i^2}{N} \leq y_{\text{max}}^2
\]

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

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

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

where \( y_{\text{max}} = \max_{y \in Y} |y| \).

**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

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

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

\[
|\alpha_t| = |y_t - \mathbf{w}_t^T \Phi(x_t)| \leq y_{\text{max}} + \| \mathbf{w}_t \| \| \Phi(x_t) \| \leq y_{\text{max}} + \| \mathbf{w}_t \|
\]

It implies that

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

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

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

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

(1)

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

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

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

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

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

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

\[
\leq \lambda^{-1} (y_{\text{max}} + s_T) = s_{T+1}
\]

\( \square \)
Theorem 8. Let us consider running of Algorithm 2 where \((x_t, y_t)\) is sampled from the training set \(D\) or the join distribution \(P_{X,Y}\). Let define the gradient error as \(M_t = \frac{\Delta^*}{\eta_t} = -f'(w^*_t, x_t, y_t)\). We have the following

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

Proof. Let us define \(\delta_t = g_t + Z_t M_t\). We have \(w_{t+1} = w_t - \eta_t \delta_t\).

\[
\|w_{t+1} - w^*\|^2 = \|w_t - \eta_t \delta_t - w^*\|^2 = \|w_t - w^*\|^2 + \eta_t^2 \|\delta_t\|^2 - 2 \eta_t \langle w_t - w^*, \delta_t \rangle
\]

Taking the conditional expectation w.r.t. \(w_t\), we achieve

\[
\langle w_t - w^*, g_t \rangle = \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2 \eta_t} + \frac{\eta_t \|\delta_t\|^2}{2} - \langle w_t - w^*, E[Z_t M_t] \rangle
\]

\[
\langle w_t - w^*, f'(w_t) \rangle = \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2 \eta_t} + \frac{\eta_t \|\delta_t\|^2}{2} - \langle w_t - w^*, E[Z_t M_t] \rangle
\]

\[
f(w_t) - f(w^*) + \frac{\lambda}{2} \|w_t - w^*\|^2 \leq \frac{\|w_t - w^*\|^2 - \|w_{t+1} - w^*\|^2}{2 \eta_t} + \frac{\eta_t \|\delta_t\|^2}{2} - \langle w_t - w^*, E[Z_t M_t] \rangle
\]

Taking expectation, we come to the following

\[
\mathbb{E}[f(w_t) - f(w^*)] \leq \frac{\lambda}{2} (t - 1) \mathbb{E}\left[\|w_t - w^*\|^2\right] - \frac{\lambda}{2} t \mathbb{E}\left[\|w_t - w^*\|^2\right] + \frac{Q}{2 \lambda T} \\
+ E\left[\|w_t - w^*\|^2\right]^{1/2} \mathbb{E}\left[\|M_t\|^2\right]^{1/2} \mathbb{E}[Z_t]^2^{1/2}
\]

Summing when \(t = 1, 2, \ldots, T\), we gain

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

Let \(w_T = \frac{1}{T} \sum_{t=1}^{T} w_t\), we reach

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

\(\square\)
Theorem 9. We denote the gap

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

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

\[ f (w_r) \leq f (w^*) + d_T + \frac{Q (\log T + 1)}{2 \lambda T} \]

Proof. Let us denote \( X = f (w_r) - f (w^*) \geq 0 \) and \( Y = \sum_{t=1}^{T} f (w_t) - f (w^*) \). Then, we have

\[ \mathbb{E}_r [X] = \mathbb{E}_r [f (w_r) - f (w^*)] = \frac{\sum_{t=1}^{T} f (w_t)}{T} - f (w^*) = Y \]

Therefore, we gain

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

or equivalently

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

where \( (x_t, y_t) \mathbb{E}_r \) 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

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

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

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

Proof. Let us define \( \delta_t = g_t - Z_t l' (w_{t'}; x_{t'}, y_{t'}) \). We have the following

\[ w_{t+1} = w_t - \eta_t \delta_t \]

\[ \| w_{t+1} - w^* \|^2 = \| w_t - \eta_t \delta_t - w^* \|^2 = \| w_t - w^* \|^2 + \eta_t^2 \| \delta_t \|^2 - 2 \eta_t \langle w_t - w^*, \delta_t \rangle \]

\[ = \| w_t - w^* \|^2 + \eta_t^2 \| \delta_t \|^2 - 2 \eta_t \langle w_t - w^*, g_t \rangle + 2 \eta_t \langle w_t - w^*, Z_t l' (w_{t'}; x_{t'}, y_{t'}) \rangle \]

Taking conditional expectation w.r.t \( w_1, x_1 \) and note that \( t' < t \), we gain

\[ \mathbb{E} \left[ \| w_{t+1} - w^* \| \right] = \mathbb{E} \left[ \| w_t - w^* \| \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \| \right] - 2 \eta_t \langle w_t - w^*, \mathbb{E} [g_t] \rangle + 2 \eta_t \langle w_t - w^*, l' (w_{t'}; x_{t'}, y_{t'}) \rangle \]

\[ \leq \mathbb{E} \left[ \| w_t - w^* \| \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \| \right] - 2 \eta_t \langle w_t - w^*, f' (w_t) \rangle + 2 \eta_t \langle w_t - w^*, p_t l' (w_{t'}; x_{t'}, y_{t'}) \rangle \]

\[ + 2 \eta_t \left( f (w_t) - f (w^*) - \frac{\lambda}{2} \| w_t - w^* \|^2 \right) \]

Since the function \( f (.) \) is \( \lambda \)-strongly convex and \( w^* \) is the optimal solution, we have

\[ f (w_t) - f (w^*) \geq \langle f' (w_t), w_t - w^* \rangle + \frac{\lambda}{2} \| w_t - w^* \|^2 \geq \frac{\lambda}{2} \| w_t - w^* \|^2 \]

7
It follows that
\[
\mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right] \leq \mathbb{E} \left[ \| w_t - w^* \|^2 \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \|^2 \right] + 2 \eta_t \langle w_t - w^*, \eta_t l(w_t^*; x_t; y_t) \rangle - 2 \eta_t \lambda \| w_t - w^* \|^2
\]

Taking expectation the above inequality, we achieve
\[
\mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right] \leq \mathbb{E} \left[ \| w_t - w^* \|^2 \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \|^2 \right] + 2 \eta_t \mathbb{E} \left[ \langle w_t - w^*, \eta_t l'(w_t^*; x_t^t; y_t^t) \rangle \right] - 2 \eta_t \lambda \| w_t - w^* \|^2
\]
\[
= \frac{t-2}{t} \mathbb{E} \left[ \| w_t - w^* \|^2 \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \|^2 \right] + 2 \eta_t \mathbb{E} \left[ \| w_t - w^* \|^2 \right]^{1/2} \mathbb{E} \left[ \| \eta_t l'(w_t^*; x_t^t; y_t^t) \|^2 \right]^{1/2} \mathbb{E} \left[ \| \delta_t \|^2 \right]^{1/2} - 2 \eta_t \lambda \| w_t - w^* \|^2
\]
\[
\leq \frac{t-2}{t} \mathbb{E} \left[ \| w_t - w^* \|^2 \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \|^2 \right] + 2 \eta_t \mathbb{E} \left[ \| w_t - w^* \|^2 \right]^{1/2} \mathbb{E} \left[ \| \eta_t l'(w_t^*; x_t^t; y_t^t) \|^2 \right]^{1/2}
\]
\[
\leq \frac{t-2}{t} \mathbb{E} \left[ \| w_t - w^* \|^2 \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \|^2 \right] + 2 \eta_t \mathbb{E} \left[ \| w_t - w^* \|^2 \right]^{1/2} \mathbb{E} \left[ \| \eta_t l'(w_t^*; x_t^t; y_t^t) \|^2 \right]^{1/2}
\]

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
\[
\mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right] \leq \frac{t-2}{t} \mathbb{E} \left[ \| w_t - w^* \|^2 \right] + \eta_t^2 \mathbb{E} \left[ \| \delta_t \|^2 \right] + 2 \eta_t \mathbb{E} \left[ \| w_t - w^* \|^2 \right]^{1/2} \mathbb{E} \left[ \| \eta_t l'(w_t^*; x_t^t; y_t^t) \|^2 \right]^{1/2}
\]

By choosing \( W_t = \frac{Q^2 \lambda^2 - M^2 / 2 \lambda^2}{t} \), we gain if \( \mathbb{E} \left[ \| w_t - w^* \|^2 \right] \leq W_t \), then \( \mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right] \leq W_{t+1} \).

**Theorem 11.** Let us consider running of Algorithm 3 where \( (x_t, y_t) \) is sampled from the training set \( D \) or the join distribution \( \mathbb{P}_{X,Y} \). Let define the gradient error as \( M_t = \frac{\Delta_t}{\eta_t} = -l'(w_t^*; x_t^t; y_t^t) \). We have the following
\[
\mathbb{E} \left[ f(w_T^*) - f(w^*) \right] \leq \frac{D^2 + Q^2 \log(1/(1-\gamma))}{2\gamma T} + \frac{D^1/2}{\gamma T} \sum_{t=(1-\gamma)T+1}^{T} \mathbb{E} \left[ \| M_t \|^2 \right]^{1/2} \tag{3}
\]

**Proof.** Let us define \( \delta_t = g_t + Z_t M_t \). We have the following
\[
\| w_{t+1} - w^* \|^2 = \| w_t - \eta_t \delta_t - w^* \|^2 = \| w_t - w^* \|^2 + \eta_t^2 \| \delta_t \|^2 - 2 \eta_t \langle w_t - w^*, \delta_t \rangle
\]
\[
\langle w_t - w^*, g_t \rangle = \frac{\| w_t - w^* \|^2 - \| w_{t+1} - w^* \|^2}{2 \eta_t} + \frac{\eta_t}{2} \| \delta_t \|^2 - \langle w_t - w^*, Z_t M_t \rangle
\]

Taking the conditional expectation w.r.t \( w_t \), we achieve
\[
\langle w_t - w^*, g_t \rangle = \frac{\mathbb{E} \left[ \| w_t - w^* \|^2 \right] - \mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right]}{2 \eta_t} + \frac{\eta_t}{2} \mathbb{E} \left[ \| \delta_t \|^2 \right] - \langle w_t - w^*, Z_t M_t \rangle
\]
\[
\langle w_t - w^*, f'(w_t) \rangle = \frac{\mathbb{E} \left[ \| w_t - w^* \|^2 \right] - \mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right]}{2 \eta_t} + \frac{\eta_t}{2} \mathbb{E} \left[ \| \delta_t \|^2 \right] - \langle w_t - w^*, Z_t M_t \rangle
\]
\[
f(w_t) - f(w^*) \leq \frac{\mathbb{E} \left[ \| w_t - w^* \|^2 \right] - \mathbb{E} \left[ \| w_{t+1} - w^* \|^2 \right]}{2 \eta_t} + \frac{\eta_t}{2} \mathbb{E} \left[ \| \delta_t \|^2 \right] - \langle w_t - w^*, Z_t M_t \rangle
\]

Taking expectation and summing when \( t = (1-\gamma) T + 1, \ldots, T \), let \( W_T^* = \frac{1}{T} \sum_{t=(1-\gamma)T+1}^{T} w_t \) and note that \( p_t \leq P(S_t = 1) \leq \frac{t}{T} \), we reach the following

\[
\mathbb{E} \left[ f(w_T^*) - f(w^*) \right] \leq \frac{D^2 + Q^2 \log(1/(1-\gamma))}{2\gamma T} + \frac{D^1/2}{\gamma T} \sum_{t=(1-\gamma)T+1}^{T} \mathbb{E} \left[ \| M_t \|^2 \right]^{1/2} \tag{3}
\]
\[
\gamma T \mathbb{E} \left[ \sum_{t=(1-\gamma)T+1}^T \frac{f(w_t) - f(w^*)}{\gamma T} \right] \leq \frac{\mathbb{E} \left[ \|w_{(1-\gamma)T+1} - w^*\|^2 \right]}{2\eta(1-\gamma)T+1} + \sum_{t=(1-\gamma)T+2}^T \mathbb{E} \left[ \|w_t - w^*\|^2 \right] \left( \frac{1}{2\eta_t} - \frac{1}{2\eta_{t-1}} \right)
\]

\begin{align*}
&\quad + \sum_{t=(1-\gamma)T+1}^T \left( \frac{\eta_t \mathbb{E} \left[ \|\delta_t\|^2 \right]}{2} + \mathbb{E} \left[ \|w_t - w^*\|^{1/2} \right] \mathbb{E} \left[ \|M_t\|^2 \right]^{1/2} \right) \\
&\quad \leq \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} \beta \frac{1}{t}
\end{align*}

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

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

Finally, we achieve

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

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

\[
\mathbb{E} [f(\overline{w}_T) - f(w^*)] \leq \frac{D\lambda^2 + Q \log (1/(1-\gamma)) + 2\beta LD^{1/2} \log (1/(1-\gamma))}{2\gamma T}
\]

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

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

\[
f(w_r) \leq f(w^*) + \frac{R}{2\gamma \delta T}
\]

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

**Proof.** Let us denote 
\[X = f(w_r) - f(w^*) \geq 0\] and 
\[Y = \frac{\sum_{t=(1-\gamma)T+1}^T f(w_t)}{\gamma T} - f(w^*) = Y\]

Then, we have

\[
\mathbb{E}_r [X] = \mathbb{E}_r [f(w_r) - f(w^*)] = \frac{\sum_{t=(1-\gamma)T+1}^T f(w_t)}{\gamma T} - f(w^*) = Y
\]
We present in detail how to incrementally maintain the inverse matrix $K_i^{-1}$. We consider two cases

- $|Z_t| \leq B$

  We compute as follows:
  
  Compute $d = K_{i-1}^{-1} k_t$
  
  Set $\|\delta_t\|^2 = K(x_t, x_t) - k_t^T d$
  
  Update $K_i^{-1} = \begin{bmatrix} K_{i-1}^{-1} & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & K_{i-1}^{-1} \end{bmatrix} + \frac{1}{\|\delta_t\|^2} \begin{bmatrix} d \\ -1 \end{bmatrix} \begin{bmatrix} d^T \\ -1 \end{bmatrix}$

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

- $|Z_t| = B + 1$

  To update $K_i^{-1}$ from $K_{i-1}^{-1}$ we observe that these two matrices $K_{i-1}$ and $K_i$ are distinct in one row and one column. Concretely, to transform $K_{i-1}$ to $K_i$, we can substitute the column $k_p$ by $k_t$ and do the same for the corresponding row. Therefore, we can formulate $K_i = K_{i-1} + L$ where $L$ is a sparse matrix of all zeros except for one column and row, which can be computed as $L_p = k_t - k_p$. It is apparent that $\text{rank}(L) = 2$.

  To update $K_i^{-1}$ from $K_{i-1}$, we rely on Thm. 14 (cf. [1]).

We assume that the $i$-th column and row in $B \times B$ matrix $K_{i-1}$ and $K_i$ is mapped to the element $x_{\pi(i)}$ in $(x_1, x_2, \ldots, x_t)$. We further assume the removal element $x_p$ locates at $m$-th column in matrix $K_{i-1}$. To gain $K_i$ from $K_{i-1}$, we replace $x_p$ by $x_t$ and hence $\pi^{-1}(t) = \pi^{-1}(p) = m$. It is evident that $K_i = K_{i-1} + L$ 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 $i = 1, \ldots, B$. It is apparent that $\text{rank}(L) = 2$ 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 $\text{rank}(L_1) = \text{rank}(L_2) = 1$.

  To directly apply Thm. 14, we denote $C_1 = A = K_{i-1}$, $B_1 = L_1$, and $B_2 = L_2$. We first compute $C_2^{-1}$ by

  \[ C_2^{-1} = C_1^{-1} - g_1 C_1^{-1} B_1 C_1^{-1} \]  

  It is obvious the computational cost to compute $C_2^{-1}$ as in Eq. (6) is $O(B^2)$.

  We then compute $K_i^{-1} = (A + B)^{-1} = (A + B_1 + B_2)^{-1}$ as

  \[ K_i^{-1} = (A + B)^{-1} = C_2^{-1} - g_2 C_2^{-1} B_2 C_2^{-1} \]  

  The computational cost of Eq. (7) is again $O(B^2)$.

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