Additional Results and Extensions for the paper "Using Taylor-Approximated Gradients to Improve the Frank-Wolfe Method for Empirical Risk Minimization"

Zikai Xiong[1] and Robert M. Freund[2]

[1]MIT Operations Research Center
[2]MIT Sloan School of Management

August 29, 2022

A *Rule−DBD*$\sqrt[4]{K}$ for Non-convex Loss Functions

In a similar spirit as *Rule−DBD*$\sqrt{k}$, we also present the deterministic rule *Rule−DBD*$\sqrt[4]{K}$ which achieves nearly identical computational guarantees (to within a constant factor) as *Rule−SBD*$\sqrt[4]{K}$. First of all, let us recall the definition of *Rule−DBD*$\sqrt[4]{K}$.

Definition A.1. *Rule−DBD*$\sqrt[4]{K}$. For a fixed value of $K \geq 1$, and for any $k \geq 1$, define:

$$\mathcal{B}_k = \begin{cases} [n] & \text{if } k/\lfloor \sqrt[4]{K} \rfloor \in \mathbb{N} \\ \emptyset & \text{if } k/\lfloor \sqrt[4]{K} \rfloor \notin \mathbb{N} \, . \end{cases}$$

In *Rule−DBD*$\sqrt[4]{K}$ we do not update any Taylor points unless k is integer times of $\lfloor \sqrt[4]{K} \rfloor$, and for these values of k we update all n Taylor points. We point out that for *Rule−DBD*$\sqrt[4]{K}$ the Taylor points are updated less often as K grows (in a different way but with similar effect as in *Rule−SBD*$\sqrt[4]{K}$). Similar to the case of *Rule−SBD*$\sqrt[4]{K}$, we have:

Proposition A.1. *Using Rule−DBD*$\sqrt[4]{K}$ *and* $K \geq 1$ *iterations, the total number of flops used in Algorithm 2.1 is* $O(K \cdot (\text{fLMO} + p^2) + K^{3/4} \cdot np^2)$.

(We omit the proof as it is nearly identical to that of Proposition 4.2.)

Theorem A.2. *Suppose that Assumption 1.1 holds and* F *is not necessarily convex, and Algorithm 2.1 with Rule−DBD*$\sqrt[4]{K}$ *is applied to the problem (1.2) with step-sizes defined by* $\gamma_k := \gamma := 1/\sqrt{K+1}$ *for all* $k \geq 0$, *where* $K \geq 1$ *is given. Then:*

$$\min_{k \in \{0,\ldots,K\}} \mathcal{G}(x^k) \leq \sum_{k=0}^{K} \frac{\mathcal{G}(x^k)}{K+1} \leq \frac{F(x^0) - F(x^\star)}{\sqrt{K+1}} + \frac{\hat{L}D_3^3 + LD_2^2}{2n\sqrt{K+1}} \, . \tag{A-1}$$

Corollary A.1. *Let*

$$K \geq \left\lceil \frac{\left(2n(F(x^0) - F(x^*)) + \hat{L}D_3^3 + LD_2^2\right)^2}{(2n\epsilon)^2} \right\rceil \, ,$$

and let the iteration index $\hat{k}$ *be chosen uniformly from* $[K]$, *namely,* $\hat{k} \sim \mathcal{U}(\{1,\ldots,K\})$. *Then* $\mathbb{E}_{\hat{k} \sim \mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$, *and the total number of flops required is at most*

$$O\left((\text{fLMO}+p^2)\left(\frac{n(F(x^0)-F(x^*))+\hat{L}D_3^3+LD_2^2}{n\epsilon}\right)^2+np^2\left(\frac{n(F(x^0)-F(x^*))+\hat{L}D_3^3+LD_2^2}{n\epsilon}\right)^{3/2}\right).$$

Also similar to Corollary 4.4, the following corollary shows that the joint dependence on n and ε is $O(n/\varepsilon^{3/2})$ under the the hypothesis that all feature vectors $w_1, w_2, \ldots$ lie in a bounded set $S \subset \{w \in \mathrm{R}^p : \|w\|_* \leq M\}$.

Corollary A.2. *Under the boundedness of the feature vectors, the bound on the number of flops in Corollary A.1 to obtain $\mathbb{E}_{\hat{k}\sim\mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$ is*

$$O\left((\text{fLMO}+p^2)\left[\frac{\left((F(x^0)-F(x^*))+LM^2\cdot\text{Diam}(\mathcal{C})^2+\hat{L}M^3\cdot\text{Diam}(\mathcal{C})^3\right)^2}{\epsilon^2}\right]\right.$$
$$\left.+np^2\left[\frac{\left((F(x^0)-F(x^*))+LM^2\cdot\text{Diam}(\mathcal{C})^2+\hat{L}M^3\cdot\text{Diam}(\mathcal{C})^3\right)^{3/2}}{\epsilon^{3/2}}\right]\right).$$

The following Table A-1 shows a comparison of the computational guarantees of the standard Frank-Wolfe method and TUFW with *Rule−SBD* $\sqrt[4]{K}$ and *Rule−DBD* $\sqrt[4]{K}$.

Table A-1: Complexity bounds for different Frank-Wolfe methods to obtain an ϵ-stationary solution of ERM_ℓ with non-convex losses, under the boundedness assumption of the feature vectors. In the table $\epsilon_0 := F(x^0) - F(x^\star)$, $c_1 := LM^2\text{Diam}(\mathcal{C})^2$, and $c_2 := \hat{L}M^3\text{Diam}(\mathcal{C})^3$.

Method	Optimality Metric	Overall Complexity
Rule−SBD $\sqrt[4]{K}$ (Cor. 4.4)	$\mathbb{E}_{\hat{k}\sim\mathcal{U}([K])}\mathbb{E}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$	$O\left((\text{fLMO}+p^2)\cdot\frac{(\epsilon_0+c_1+c_2)^2}{\epsilon^2}+np^2\cdot\frac{(\epsilon_0+c_1+c_2)^{3/2}}{\epsilon^{3/2}}\right)$
Rule−DBD $\sqrt[4]{K}$ (Cor. A.2)	$\mathbb{E}_{\hat{k}\sim\mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$	$O\left((\text{fLMO}+p^2)\cdot\frac{(\epsilon_0+c_1+c_2)^2}{\epsilon^2}+np^2\cdot\frac{(\epsilon_0+c_1+c_2)^{3/2}}{\epsilon^{3/2}}\right)$
Standard Frank-Wolfe	$\mathbb{E}_{\hat{k}\sim\mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$	$O\left((\text{fLMO}+np)\cdot\frac{(\epsilon_0+c_1)^2}{\epsilon^2}\right)$

A.1 Proof of Theorem A.2

Let us first prove the following lemma.

Lemma A.1. *In any iteration k of Algorithm 2.1 with Rule−DBD* $\sqrt[4]{K}$ *and $\gamma_k := \gamma := 1/\sqrt{K+1}$, for problem setup (1.2), it holds that*

$$\sum_{i=1}^{n}\left(\sum_{j=\tau_i^k}^{k-1}\left|w_i^\top(z^j-y^j)\right|\right)^2\cdot|w_i^\top(z^k-\bar{z}^k)| \leq K^{1/2}D_3^3 \tag{A-2}$$

holds for any $y^j, z^j, \bar{z}^j \in \mathcal{C}$, $j = 1, \ldots, k$.

Proof of Lemma A.1. Notice in *Rule−DBD* $\sqrt[4]{K}$ that $k-\tau_i^k \leq \lfloor K^{1/4}\rfloor - 1$. Now let $u, v \geq 1$ satisfy $\frac{1}{u}+\frac{1}{v}=1$.

Then it holds that

$$\begin{aligned}
&\sum_{i=1}^{n}\Big(\sum_{j=\tau_i^k}^{k-1}\big|w_i^\top(z^j-y^j)\big|\Big)^2\cdot|w_i^\top(z^k-\bar{z}^k)| \\
\leq&(\lfloor K^{1/4}\rfloor-1)\sum_{i=1}^{n}\sum_{j=\max\{0,k+1-\lfloor K^{1/4}\rfloor\}}^{k-1}\big|w_i^\top(z^j-y^j)\big|^2\cdot|w_i^\top(z^k-\bar{z}^k)| \qquad \text{(A-3)}\\
\leq&(\lfloor K^{1/4}\rfloor-1)\sum_{j=\max\{0,k+1-\lfloor K^{1/4}\rfloor\}}^{k-1}D_{2u}^2D_v\leq(\lfloor K^{1/4}\rfloor-1)^2D_{2u}^2D_v\ ,
\end{aligned}$$

where the first inequality uses $\|a\|_1^2\leq d\|a\|_2^2$ for $a\in\mathbb{R}^d$, and the second inequality follows similar to the equivalent part of proof of Lemma 3.20 that uses $\sum_{i=1}^n a_i^2b_i\leq\|a\|_{2u}^2\|b\|_v$ for $a,b\in\mathbb{R}^n$ and $u,v\geq1$ and $\frac{1}{u}+\frac{1}{v}=1$. Setting $u=\frac{3}{2}$ and $v=3$, the right-hand side of (A-3) is bounded above by $K^{1/2}D_3^3$. □

And then we can prove Theorem A.2.

Proof of Theorem A.2. The first inequality in (A-1) is obvious. For the second inequality, note from Lemma 4.6 that:

$$\sum_{k=0}^{K}\frac{\mathcal{G}(x^k)}{K+1}\leq\frac{2n\epsilon_0+LD_2^2}{2n\sqrt{K+1}}+\frac{1}{K+1}\sum_{k=0}^{K}(\nabla F(x^k)-g^k)^\top(s^k-\bar{s}^k)\ . \qquad \text{(A-4)}$$

Let Q_K denote the second term of the right-hand side of (A-4). We have

$$Q_K\leq\frac{\hat{L}}{2n(K+1)^2}\sum_{k=0}^{K}\sum_{i=1}^{n}\Big(\sum_{j=\tau_i^k}^{k-1}\big|w_i^\top(s^j-x^j)\big|\Big)^2\cdot|w_i^\top(s^k-\bar{s}^k)|\leq\frac{\hat{L}D_3^3}{2n\sqrt{K+1}}\ , \qquad \text{(A-5)}$$

where the first inequality is due to Lemma 3.14, and the second inequality is due to Lemma A.1. Substituting this bound back to (A-4) yields the second inequality of (A-1). □

B Adaptive Step-size

In this section we are going to introduce the adaptive-step size proposed in (5.1) and prove the worst-case convergence rates in the case of using the TUFW with *Rule−SBD*$\sqrt{k}$ on (1.2) with convex objectives. Other rules are similar and less complicated.

We first recall the adaptive step-size as follows:

$$\tilde{\gamma}_k:=\begin{cases}\min\Big\{\gamma_k,\frac{(g^k)^\top(x^k-s^k)}{(s^k-x^k)^\top H_k(s^k-x^k)}\Big\} & \text{when } (s^k-x^k)^\top H_k(s^k-x^k)>0\ ,\\ \gamma_k & \text{when } (s^k-x^k)^\top H_k(s^k-x^k)\leq0\ ,\end{cases} \qquad \text{(B-6)}$$

where H_k is defined in (2.1) and γ_k is the standard step-size, which is $\frac{2}{k+2}$ for convex loss functions and $\frac{1}{\sqrt{K+1}}$ for non-convex loss functions.

This adaptive step-size $\tilde{\gamma}_k$ can approximately minimize the quadratic approximation of the objective function in the range of $[0,\gamma_k]$. Let $x(\lambda):=x^k+\gamma(s^k-x^k)$ and then

$$\begin{aligned}
&F(x(\gamma))=F(x^k)+\gamma(g^k)^\top(s^k-x^k)+\frac{\gamma^2}{2}(s^k-x^k)^\top H_k(s^k-x^k)\\
&+\frac{1}{n}\sum_{i=1}^{n}\int_{t=0}^{\gamma}\Big(\nabla f_i(x^k+t(s^k-x^k))-\nabla f_i(b_i)-\nabla^2 f_i(b_i)\big(x^k+t(s^k-x^k)-b_i\big)\Big)^\top(s^k-x^k)\mathrm{d}t.
\end{aligned} \qquad \text{(B-7)}$$

It could be further proven that when $\gamma \in [0, \gamma_k]$, the first three terms of the right-hand side dominates and therefore $\tilde{\gamma}_k$ defined in (5.1), which is also the closed-form solution of

$$\arg\min_{\gamma \in [0,\gamma_k]} F(x^k) + \gamma (g^k)^\top (s^k - x^k) + \frac{\gamma^2}{2} (s^k - x^k)^\top H_k (s^k - x^k) \ ,$$

can be approximately regarded as $\arg\min_{\gamma \in [0,\gamma_k]} F(x(\gamma))$, which yields more decrease of the objective value than the standard step-size. Then we have the following theorem.

Theorem B.1. *Suppose that F is convex and Assumption 1.1 holds, and Algorithm 2.1 with Rule$-$SBD$\sqrt{k}$ is applied to the problem (1.2) with adaptive step-sizes defined by (B-6) for all $k \geq 0$. Then for all $k \geq 1$ we have:*

$$\mathbb{E}[F(x^k) - F(x^\star)] \leq \frac{2LD_2^2 + 410\hat{L}D_1D_\infty^2}{n(k+1)} \ . \tag{B-8}$$

Proof of Theorem B.1. First of all, suppose that the $\{\tilde{x}^l\}_l$ are the iterates of the TUFW with adaptive step-sizes and $\{s^l\}_l$ are still the outputs of the linear minimization oracle on $\{\tilde{x}^l\}_l$. We define $\delta_k := F(\tilde{x}^{k+1}) - F(\tilde{x}^k + \gamma_k(s^k - \tilde{x}^k))$, the difference of using adaptive step-sizes and standard step-sizes. Similar with the proof of Lemma 3.17, we have

$$\begin{aligned} F\big(\tilde{x}^{k+1}\big) = \ & F\big(\tilde{x}^k + \gamma_k(s^k - \tilde{x}^k)\big) + \delta_k \\ \leq \ & F(\tilde{x}^k) + \gamma_k \langle \nabla F(\tilde{x}^k) - g^k, s^k - x^\star \rangle + \gamma_k (F(x^\star) - F(\tilde{x}^k)) + \gamma_k^2 LD_2^2/2n + \delta_k \ , \end{aligned}$$

where the inequality is due to Lemma 3.12. Subtracting $F^\star$ from both sides of the above inequality chain, we arrive at:

$$\tilde{\epsilon}_{k+1} \leq (1 - \gamma_k)\tilde{\epsilon}_k + \gamma_k \left(\nabla F(\tilde{x}^k) - g^k\right)^\top (s^k - x^\star) + \gamma_k^2 LD_2^2/2n + \delta_k \ ,$$

where $\tilde{\epsilon}_k$ denotes $F(\tilde{x}^k) - F(x^\star)$. Multiplying both side by $(k+1)(k+2)$ and telescoping the inequalities yields:

$$\begin{aligned} (k+1)(k+2)\tilde{\epsilon}_{k+1} \leq & \frac{2(k+1)LD_2^2}{n} + \sum_{t=1}^{k} 2(t+1)(\nabla F(\tilde{x}^t) - g^t)^\top (s^t - x^\star) \\ & + \sum_{t=0}^{k} (t+1)(t+2)\delta_t \ . \end{aligned} \tag{B-9}$$

Now it is time to study the upper bound of δ_k. According to (B-7), we can write the δ_k as follows

$$\begin{aligned} \delta_k = & \Big(F(x^k) + \tilde{\gamma}_k (g^k)^\top (s^k - x^k) + \frac{\tilde{\gamma}_k}{2} (s^k - x^k)^\top H_k (s^k - x^k)\Big) \\ & - \Big(F(x^k) + \gamma_k (g^k)^\top (s^k - x^k) + \frac{\gamma_k^2}{2} (s^k - x^k)^\top H_k (s^k - x^k)\Big) \\ & + \frac{1}{n} \sum_{i=1}^{n} \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} \Big(\nabla f_i(x^k + \alpha(s^k - x^k)) - \nabla f_i(b_i) - \nabla^2 f_i(b_i)\big(x^k + \alpha(s^k - x^k) - b_i\big)\Big)^\top (s^k - x^k) \mathrm{d}\alpha \end{aligned} \tag{B-10}$$

where

$$\begin{aligned} & F(x^k) + \tilde{\gamma}_k (g^k)^\top (s^k - x^k) + \frac{\tilde{\gamma}_k}{2} (s^k - x^k)^\top H_k (s^k - x^k) \leq \\ & F(x^k) + \gamma_k (g^k)^\top (s^k - x^k) + \frac{\gamma_k^2}{2} (s^k - x^k)^\top H_k (s^k - x^k) \end{aligned}$$

because of the definition of adaptive step-sizes in (B-6). Now

$$\begin{aligned} & \delta_k \leq \\ & \frac{1}{n} \sum_{i=1}^{n} \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} \Big(\nabla f_i(x^k + \alpha(s^k - x^k)) - \nabla f_i(b_i) - \nabla^2 f_i(b_i)\big(x^k + \alpha(s^k - x^k) - b_i\big)\Big)^\top (s^k - x^k) \mathrm{d}\alpha \end{aligned} \tag{B-11}$$

For simplicity of notations, we use $C_i(\alpha)$ to denote the component inside the i-th integral of the right-hand side of (B-11), which is

$$\begin{aligned} C_i(\alpha) := \Big(& l_i'\big(w_i^\top \tilde{x}^k + \alpha(w_i^\top s^k - w_i^\top \tilde{x}^k)\big) - l_i'(w_i^\top \tilde{x}^{\tau_i^k}) \\ & - l_i''(w_i^\top x^{\tau_i^k})\big(w_i^\top \tilde{x}^k + \alpha(w_i^\top s^k - w_i^\top \tilde{x}^k) - w_i^\top \tilde{x}^{\tau_i^k}\big)\Big) w_i^\top (s^k - \tilde{x}^k) \ , \end{aligned}$$

and then for each $i = 1, \ldots, n$, due to Assumption 1.4

$$\begin{aligned} C_i(\alpha) \leq & \tfrac{\hat{L}}{2} \left| w_i^\top \tilde{x}^k + \alpha(w_i^\top s^k - w_i^\top \tilde{x}^k) - w_i^\top \tilde{x}^{\tau_i^k} \right|^2 \left| w_i^\top (s^k - \tilde{x}^k) \right| \\ & \leq \hat{L} \left| w_i^\top \tilde{x}^k - w_i^\top \tilde{x}^{\tau_i^k} \right|^2 \left| w_i^\top (s^k - \tilde{x}^k) \right| + \hat{L}\alpha^2 \left| w_i^\top (s^k - \tilde{x}^k) \right|^3 \ . \end{aligned} \tag{B-12}$$

Now, for any $k \geq 0$

$$\begin{aligned} \delta_k \leq & \sum_{i=1}^n \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} C_i(\alpha) \mathrm{d}\alpha \leq \frac{1}{n} \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} \sum_{i=1}^n |C_i(\alpha)| \mathrm{d}\alpha \\ & \leq \frac{\hat{L}}{n} \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} \sum_{i=1}^n \left| w_i^\top \tilde{x}^k - w_i^\top \tilde{x}^{\tau_i^k} \right|^2 \left| w_i^\top (s^k - \tilde{x}^k) \right| \mathrm{d}\alpha + \frac{\hat{L}}{n} \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} \sum_{i=1}^n \alpha^2 \left| w_i^\top (s^k - \tilde{x}^k) \right|^3 \mathrm{d}\alpha \\ & \leq \frac{\hat{L}\gamma_k}{n} \sum_{i=1}^n \left| w_i^\top \tilde{x}^k - w_i^\top \tilde{x}^{\tau_i^k} \right|^2 \left| w_i^\top (s^k - \tilde{x}^k) \right| + \frac{\hat{L}\gamma_k^3}{n} \sum_{i=1}^n \left| w_i^\top (s^k - \tilde{x}^k) \right|^3 \ , \end{aligned}$$

where the third inequality is due to (B-12) and the fourth inequality is due to $\tilde{\gamma}_k \leq \gamma_k$. Furthermore,

$$\frac{\hat{L}\gamma_k^3}{n} \sum_{i=1}^n \left| w_i^\top (s^k - \tilde{x}^k) \right|^3 \leq \frac{\hat{L}\gamma_k^3}{n} D_3^3 \tag{B-13}$$

and we can still use the inequalities in (3.8) from Lemma 3.14 to obtain

$$\begin{aligned} & \frac{\hat{L}\gamma_k}{n} \sum_{i=1}^n \left| w_i^\top \tilde{x}^k - w_i^\top \tilde{x}^{\tau_i^k} \right|^2 \left| w_i^\top (s^k - \tilde{x}^k) \right| \\ & \leq \frac{\hat{L}\gamma_k}{n} \sum_{i=1}^n \Big(\sum_{j=\tau_i^k}^{k-1} \left| \tilde{\gamma}_j w_i^\top (s^j - \tilde{x}^j) \right| \Big)^2 \cdot |w_i^\top (s^k - \tilde{x}^k)| \\ & \leq \frac{\gamma_k \hat{L} D_\infty}{n} \sum_{i=1}^n \Big(\sum_{j=\tau_i^k}^{k-1} \left| \tilde{\gamma}_j w_i^\top (s^j - \tilde{x}^j) \right| \Big)^2 \ . \end{aligned} \tag{B-14}$$

In general, for any $t \geq 0$,

$$\delta_t \leq \frac{\gamma_t \hat{L} D_\infty}{n} \sum_{i=1}^n \Big(\sum_{j=\tau_i^t}^{t-1} \left| \tilde{\gamma}_j w_i^\top (s^j - \tilde{x}^j) \right| \Big)^2 + \frac{\hat{L}\gamma_t^3}{n} D_3^3 \ . \tag{B-15}$$

In addition, directly using Lemma 3.14 yields

$$\begin{aligned} \left(\nabla F(\tilde{x}^t) - g^t \right)^\top (s^t - x^\star) \leq & \frac{\hat{L}}{2n} \sum_{i=1}^n \Big(\sum_{j=\tau_i^t}^{t-1} \tilde{\gamma}_j \left| w_i^\top (s^j - \tilde{x}^j) \right| \Big)^2 \cdot |w_i^\top (s^t - x^\star)| \\ \leq & \frac{\hat{L} D_\infty}{2n} \sum_{i=1}^n \Big(\sum_{j=\tau_i^t}^{t-1} \tilde{\gamma}_j \left| w_i^\top (s^j - \tilde{x}^j) \right| \Big)^2 \ . \end{aligned} \tag{B-16}$$

for any $t \geq 1$. Substituting (B-15), (B-16), and $\gamma_t := \frac{2}{t+2}$ into (B-9) yields

$$(k+1)(k+2)\tilde{\epsilon}_{k+1} \leq \frac{2(k+1)LD_2^2}{n} + \sum_{t=1}^{k} 3(t+1)\frac{\hat{L}D_\infty}{n} \sum_{i=1}^{n} \Big(\sum_{j=\tau_i^t}^{t-1} \big|\tilde{\gamma}_j w_i^\top (s^j - \tilde{x}^j)\big| \Big)^2 + \sum_{t=0}^{k} \frac{8\hat{L}}{(t+2)n} D_3^3 \,. \tag{B-17}$$

Since $\tilde{\gamma}_t \leq \frac{2}{t+2}$ for any $t \geq 0$, using Lemma 3.18 yields

$$\mathbb{E}\left[\sum_{i=1}^{n} \Big(\sum_{j=\tau_i^t}^{t-1} \big|\tilde{\gamma}_j w_i^\top (s^j - \tilde{x}^j)\big| \Big)^2 \right] \leq \frac{134 D_1 D_\infty}{t+2} \,.$$

Since $\sum_{t=0}^{k} \frac{1}{t+2} \leq \log(k+1)$, applying expectation on both sides of (B-17) yields

$$\begin{aligned} (k+1)(k+2)\mathbb{E}\tilde{\epsilon}_{k+1} \leq & \frac{2(k+1)LD_2^2}{n} + \sum_{t=0}^{k} 402 \frac{\hat{L}D_1 D_\infty^2}{n} + \log(k+1) 8\hat{L}D_3^3/n \\ \leq & \frac{2(k+1)LD_2^2}{n} + 410(k+1)\frac{\hat{L}D_1 D_\infty^2}{n} \,, \end{aligned}$$

where the second inequality is due to $D_3^3 \leq D_1 D_\infty^2$ and $\log(k+1) \leq k+1$ when $k \geq 0$. Now this inequality above can directly lead to (B-8).

□

C Extension to General ERM Problems

In this section, we extend TUFW to the general ERM problems. Instead of Assumption 1.1, the new assumption for (1.1) is as follows:

Assumption C.1. *The following conditions hold for* (1.1).

1. *The feasibility set $\mathcal{C}$ is a compact convex set with diameter*

$$\mathrm{Diam}(\mathcal{C}) = \max_{u,v \in \mathcal{C}} \|u - v\| \tag{C-18}$$

and the linear minimization problem $\arg\min_{x \in \mathcal{C}} g^\top x$ can be easily solved for any $g \in \mathbb{R}^p$,

2. *for the multivariate loss function $f_i(\cdot)$, $i = 1, \ldots, n$, in (1.2), the gradient $\nabla f_i(\cdot)$ is L-Lipschitz continuous on $\mathcal{C}$, namely*

$$\|\nabla f_i(u) - \nabla f_i(v)\|_* \leq L\|u - v\|, \quad \text{for any } u,\ v \in \mathcal{C}, \tag{C-19}$$

where $\|\cdot\|_*$ is the dual norm of the norm on the space of variables,

3. *and the Hessian matrix $\nabla^2 f_i(\cdot)$ is $\hat{L}$-Lipschitz continuous on $\mathcal{C}$, namely*

$$\|\nabla^2 f_i(u) - \nabla^2 f_i(v)\| \leq \hat{L}\|u - v\|, \quad \text{for any } u,\ v \in \mathcal{C}, \tag{C-20}$$

where the norm of Hessian matrices is the operator norm induced by the norm on the space of variables.

In the rest of this section, we will use $x^\star$ to denote the optimal solution of problem (1.1) and show how to prove the corresponding new convergence rates. First of all, we have the following fact.

Fact C.1. *Under the continuity (C-19) and (C-20) defined in Assumption C.1, for any x, y in $\mathcal{C}$ and $i \in [n]$, it holds that*

$$|f_i(y) - f_i(x) - \langle \nabla f_i(x), y - x\rangle| \leq L\|x - y\|^2/2 \tag{C-21}$$

and

$$\|\nabla f_i(y) - \nabla f_i(x) - \langle \nabla^2 f_i(x), y - x\rangle\|_* \leq \hat{L}\|x - y\|^2/2. \tag{C-22}$$

Now we can establish a fundamental lemma that is similar to Lemma 3.14, useful in measuring the error of Taylor-estimated gradients.

Lemma C.1. *Under Assumption C.1 for the problem ERM, for any $u, v \in \mathcal{C}$,*

$$\left(\nabla F(x^k) - g^k\right)^\top (u - v) \leq \frac{\hat{L}D^2}{2n} \sum_{i=1}^{n} \Big(\sum_{j=\tau_i^k}^{k-1} \gamma_j \Big)^2 \tag{C-23}$$

holds for iteration k of Algorithm 2.1 for all $k \geq 1$.

Proof of Lemma C.1. First of all, $\left(\nabla F(x^k) - g^k\right)^\top (u - v)$ can be rewritten as

$$\frac{1}{n} \sum_{i=1}^{n} \Big(\nabla f_i\big(x^k\big) - \nabla f_i\big(x^{\tau_i^k}\big) - \nabla^2 f_i\big(x^{\tau_i^k}\big)\big(x^k - x^{\tau_i^k}\big) \Big)^\top (u - v), \tag{C-24}$$

which, since $\|u - v\| \leq D$, is smaller than or equal to

$$\frac{D}{n} \sum_{i=1}^{n} \Big\| \nabla f_i\big(x^k\big) - \nabla f_i\big(x^{\tau_i^k}\big) - \nabla^2 f_i\big(x^{\tau_i^k}\big)\big(x^k - x^{\tau_i^k}\big) \Big\|_* \leq \frac{D\hat{L}}{2n} \sum_{i=1}^{n} \big\| x^k - x^{\tau_i^k} \big\|^2, \tag{C-25}$$

where the inequality is due to (C-22) in Fact C.1. Besides, $\|x^k - x^{\tau_i^k}\|^2 \leq \big(\sum_{j=\tau_i^k}^{k-1} \|x^{j+1} - x^j\|\big)^2 \leq \big(\sum_{j=\tau_i^k}^{k-1} \gamma_j D\big)^2$. Substituting it into (C-25) yields (C-23). □

Now we first analyze the convergence of problems with convex objectives. Similar with Lemma 3.17 for problem setup (1.2), we have the following lemma for problem setup (1.1).

Lemma C.2. *When F is convex and Assumption C.1 holds, then for any k,*

$$F(x^k) - F(x^\star) \leq \frac{2LD^2}{(k+1)} + \frac{2}{(k+1)(k+2)} \sum_{t=1}^{k} t \left(\nabla F(x^{t-1}) - g^{t-1}\right)^\top (s^{t-1} - x^\star), \tag{C-26}$$

for problem setup (1.1) with $\gamma_k := 2/(k+2)$.

The proof is almost the same with the proof of Lemma 3.17, different by replacing all $\frac{D_2^2}{n}$ by D^2 and replacing Lemma 3.13 by Fact C.1.

We will then show how to modify the proofs in section 3 and section 4 to prove the convergence of general ERM problems.

Taking the TUFW with *Rule−SBD*$\sqrt{k}$ and standard step-size as an example, we have the following theorem.

Theorem C.1. *Suppose that F is convex and Assumption C.1 holds, and Algorithm 2.1 with Rule−SBD*$\sqrt{k}$ *is applied to the problem (1.1) with step-sizes defined by $\gamma_k := 2/(k+2)$ for all $k \geq 0$. Then for all $k \geq 1$ we have:*

$$\mathbb{E}[F(x^k) - F(x^\star)] \leq \frac{2LD^2 + 134\hat{L}D^3}{k+1} \ . \tag{C-27}$$

The proof is almost the same with the proof of Theorem 3.3, except replacing Lemma 3.17 and Remark 3.15 by Lemmas C.2 and C.1.

As for applying *Rule−DBD*$\sqrt{k}$ and adaptive steps, the convergence results and corresponding proofs are similar with those for problem setup (1.2) in section 3.

When the objective function is nonconvex, similar with Lemma 4.6 for problem setup (1.2), we have the following lemma for problem setup (1.1).

Lemma C.3. *Under Assumption C.1 for (1.1), if we use the step-size* $\gamma_k := \gamma := \frac{1}{\sqrt{K+1}}$ *in Algorithm 2.1, then*

$$\sum_{k=0}^{K} \frac{\mathcal{G}(x^k)}{K+1} \leq \frac{F(x^0) - F(x^\star)}{\sqrt{K+1}} + \frac{1}{K+1} \sum_{k=0}^{K} (\nabla F(x^k) - g^k)^\top (s^k - \bar{s}^k) + \frac{LD^2}{2\sqrt{K+1}}. \tag{C-28}$$

The proof is almost the same with the proof of Lemma 4.6, except replacing all $\frac{D_2^2}{n}$ by D^2 and replacing Lemma 3.13 by Fact C.1.

Taking the TUFW with *Rule−SBD*$\sqrt[4]{K}$ and standard step-size as an example, we have the following theorem.

Theorem C.2. *Suppose that Assumption C.1 holds and F is not necessarily convex, and Algorithm 2.1 with Rule−SBD*$\sqrt[4]{K}$ *is applied to the problem (1.1) with step-sizes defined by* $\gamma_k := \gamma := 1/\sqrt{K+1}$ *for all* $k \geq 0$*, where* $K \geq 1$ *is given. Then:*

$$\mathbb{E}[\min_{k \in \{0,\ldots,K\}} \mathcal{G}(x^k)] \leq \sum_{k=0}^{K} \frac{\mathbb{E}[\mathcal{G}(x^k)]}{K+1} \leq \frac{F(x^0) - F(x^*)}{\sqrt{K+1}} + \frac{3\hat{L}D^3 + LD^2}{2\sqrt{K+1}} \ . \tag{C-29}$$

The proof is almost the same with the proof of Theorem 4.3, different by replacing Lemma 4.6 and Remark 3.15 by Lemmas C.3 and C.1.

Finally, the overall complexiy result is as follows:

Corollary C.1. *Suppose that the above assumption holds, then*

1. *when F is convex, and* $\gamma_k := \frac{2}{k+2}$*, and Algorithm 2.1 with Rule−SBD*$\sqrt{k}$ *(or Rule−DBD*$\sqrt{k}$*) is applied to the problem (1.1), in order to obtain* $\mathbb{E}[F(x^k) - F(x^\star)] \leq \epsilon$ *(or* $F(x^k) - F(x^\star) \leq \epsilon$*), the bound of the number of flops is*

$$O\left((\text{fLMO} + p^2) \cdot \frac{L \cdot \text{Diam}(\mathcal{C})^2 + \hat{L} \cdot \text{Diam}(\mathcal{C})^3}{\varepsilon} + np^2 \cdot \frac{\sqrt{L \cdot \text{Diam}(\mathcal{C})^2 + \hat{L} \cdot \text{Diam}(\mathcal{C})^3}}{\sqrt{\varepsilon}} \right),$$

2. *and if* $\gamma_k := \frac{1}{\sqrt{K+1}}$*, and Algorithm 2.1 with Rule−SBD*$\sqrt[4]{K}$ *(or Rule−DBD*$\sqrt[4]{K}$*) is applied to the problem (1.1), in order to obtain* $\mathbb{E}_{\hat{k} \sim \mathcal{U}([K])} \mathbb{E}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$ *(or* $\mathbb{E}_{\hat{k} \sim \mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$ *), the bound of the number of flops is*

$$O\left((\text{fLMO} + p^2) \left[\frac{\left((F(x^0) - F(x^*)) + L \cdot \text{Diam}(\mathcal{C})^2 + \hat{L} \cdot \text{Diam}(\mathcal{C})^3 \right)^2}{\epsilon^2} \right] \right.$$
$$\left. + np^2 \left[\frac{\left((F(x^0) - F(x^*)) + L \cdot \text{Diam}(\mathcal{C})^2 + \hat{L} \cdot \text{Diam}(\mathcal{C})^3 \right)^{3/2}}{\epsilon^{3/2}} \right] \right) \ .$$

D More experimental results

In order to test our TUFW methods on problems with larger feasible regions, we increased the size of the feasibility set in (6.1) by inflating the value of λ to $\lambda' = 100\lambda$ (where recall λ was determined by cross validation). Table D-2 shows the results of these experiments. For these problems with larger feasible regions, the advantage of the TUFW methods is even more pronounced. Curiously, this increased advantage of TUFW due to a larger feasible region is not indicated by any of the theory we developed. TUFW methods and FW-ada exhibit linear-like convergence rates, but TUFW methods require far lower CPU runtime than all other methods.

Table D-2: Comparison of average CPU runtimes (in seconds) required to achieve $\mathcal{G}(x^k) \leq \epsilon$ for methods on the logistic regression problem (6.1) with λ inflated to $\lambda' = 100\lambda$. (A blank indicates the method used more than 5000 seconds.)

ϵ	dataset	n	p	*Rule−SBD*$\sqrt{k}$	*Rule−DBD*$\sqrt{k}$	FW	FW-ada	SPIDER-FW	CSFW	Speed-up
1e0	a1a	1605	123	1.08	**0.56**	3322.66	21.88			**39.14**
1e-2	a1a	1605	123	61.47	**22.15**		2671.01			**120.59**
1e-4	a1a	1605	123	4561.28	**1683.57**					
1e0	a2a	2265	123	**2.40**	4.30	3858.43	32.69			**13.64**
1e-2	a2a	2265	123	6.70	**5.68**		1959.43			**345.13**
1e-4	a2a	2265	123	28.72	**13.29**					
1e0	a8a	22696	123	15.35	**8.50**		268.54			**31.59**
1e-2	a8a	22696	123	34.86	**17.75**					
1e-4	a8a	22696	123	60.72	**30.10**					
1e0	a9a	32561	123	20.80	**10.97**		326.83			**29.79**
1e-2	a9a	32561	123	52.70	**24.35**					
1e-4	a9a	32561	123	97.60	**45.91**					
1e0	w1a	2477	300	16.17	**8.94**		4569.30			**511.10**
1e-2	w1a	2477	300	75.05	**34.01**					
1e-4	w1a	2477	300	1687.82	**560.34**					
1e0	w2a	3470	300	34.13	**18.39**					
1e-2	w2a	3470	300	138.82	**65.27**					
1e-4	w2a	3470	300	2311.84	**778.48**					
1e0	w7a	24692	300	147.81	**123.91**					
1e-2	w7a	24692	300	443.15	**339.27**					
1e-4	w7a	24692	300	3434.91	**1740.40**					
1e0	w8a	49749	300	522.63	**165.85**					
1e-2	w8a	49749	300	1053.55	**515.06**					
1e-4	w8a	49749	300		**4608.20**					
1e-1	svmguide3	1243	22	3.58	**1.17**		127.60			**109.24**
1e-3	svmguide3	1243	22	11.28	**3.67**		485.90			**132.22**
1e-5	svmguide3	1243	22	18.83	**6.08**		844.46			**138.80**
1e-7	svmguide3	1243	22	26.37	**8.54**		1201.28			**140.65**
1e-1	phishing	11055	68	2.26	**1.20**	66.23	254.91	3563.61	64.96	**53.93**
1e-3	phishing	11055	68	5.15	**2.45**	3958.88	4057.22			**1613.39**
1e-5	phishing	11055	68	19.12	**8.35**					
1e-7	phishing	11055	68	592.44	**207.94**					
1e-1	ijcnn1	49990	22	1.52	**0.47**		100.76			**213.41**
1e-3	ijcnn1	49990	22	2.32	**0.64**		243.63			**378.96**
1e-5	ijcnn1	49990	22	2.92	**0.80**		425.17			**531.71**
1e-7	ijcnn1	49990	22	3.45	**0.92**		607.30			**660.82**
1e-1	covtype	581012	54	250.33	**115.22**					
1e-3	covtype	581012	54	1163.77	**484.39**					
1e-5	covtype	581012	54	2230.09	**890.50**					

Table D-3 is almost identical to Table 4. The only difference is that the numbers in parentheses in the table are the number of iterations K at which the given average Frank-Wolfe gap was attained.

Table D-3: Comparison of average CPU runtimes (in seconds) required to achieve $\frac{1}{K+1}\sum_{k=0}^{K}\mathcal{G}(x^k) \leq \epsilon$ for methods on the non-convex binary classification problem (6.2). The numbers in parentheses are the number of iterations K at which the given average Frank-Wolfe gap was attained. (A blank indicates the method used more than 5000 seconds.)

$\mathcal{G}(x^k)$	dataset	n	p	*Rule−SBD* $\sqrt[4]{K}$	*Rule−DBD* $\sqrt[4]{K}$	FW	FW-ada	SPIDER-FW	CASPIDERG	Speed-up
1e-2	a1a	1605	119	6.05(1.8e4)	**4.44(2.3e4)**	10.46(3.7e6)	9.00(5.3e6)	15.77(2.5e7)		**2.03**
1e-3	a1a	1605	119	90.34(1.6e5)	**65.11(1.6e5)**	818.40(4.2e7)	237.93(9.2e6)			**3.65**
1e-4	a1a	1605	119	2225.61(6.3e6)	**1453.87(6.3e6)**		4953.53(2.5e7)			**3.41**
1e-2	a2a	2265	119	6.47(2.3e4)	**4.94(1.6e4)**	12.62(5.2e6)	10.55(2.1e7)	13.54(2.1e7)		**2.14**
1e-3	a2a	2265	119	93.17(2.0e5)	**70.69(2.0e5)**	781.05(4.2e7)	303.67(2.1e7)			**4.30**
1e-4	a2a	2265	119	2168.72(6.3e6)	**1389.71(9.4e6)**					
1e-2	a8a	22696	123	46.68(2.5e4)	41.92(2.5e4)	243.81(8.9e6)	140.10(8.7e6)	**35.70(4.2e7)**		0.85
1e-3	a8a	22696	123	668.35(1.8e5)	**603.88(2.5e5)**		3317.18(2.6e6)			**5.49**
1e-4	a8a	22696	123							
1e-2	a9a	32561	123	83.24(1.4e4)	79.91(1.2e4)	358.08(6.3e6)	240.37(2.6e6)	**46.05(4.2e7)**		0.58
1e-3	a9a	32561	123	1165.35(1.3e5)	**1106.16(1.3e5)**					
1e-4	a9a	32561	123							
5e-2	w1a	2477	300	8.99(2.5e4)	8.82(2.9e4)	**0.44(1.8e4)**	12.66(1.2e6)	0.96(6.6e5)	101.83(4.2e7)	0.05
1e-2	w1a	2477	300	74.94(3.3e5)	102.18(5.2e5)	**4.69(8.2e4)**	157.90(5.1e6)	19.96(1.3e7)		0.06
2e-3	w1a	2477	300	1278.96(1.5e7)	4063.80(4.2e7)	**109.23(1.4e6)**	1768.32(1.2e7)			0.09
5e-2	w2a	3470	300	12.64(1.5e4)	11.90(1.0e4)	**0.50(1.0e4)**	17.41(7.0e6)	1.01(7.9e5)	120.51(3.8e7)	0.04
1e-2	w2a	3470	300	93.47(2.8e5)	122.01(2.9e5)	**7.44(8.2e4)**	226.70(8.0e6)	21.08(1.7e7)		0.08
2e-3	w2a	3470	300	900.09(3.1e6)	2545.77(4.2e7)	**152.15(2.4e6)**	2415.59(1.4e7)			0.17
5e-2	w7a	24692	300	95.86(1.4e4)	90.12(1.6e4)	7.79(1.2e4)	271.26(2.3e6)	**2.16(6.6e5)**	595.38(2.5e7)	0.02
1e-2	w7a	24692	300	544.57(1.6e5)	561.90(4.9e4)	80.96(4.9e4)	3596.39(2.0e7)	**29.75(8.4e6)**		0.05
2e-3	w7a	24692	300	4078.15(1.0e6)		**1631.47(6.6e5)**		2990.71(4.2e7)		0.40
5e-2	w8a	49749	300	222.71(2.3e4)	216.21(1.0e4)	16.41(1.4e4)	534.20(8.5e6)	**3.25(5.2e5)**	1122.68(2.7e7)	0.02
1e-2	w8a	49749	300	1292.84(4.1e4)	1303.19(4.1e4)	176.70(9.8e4)		**41.53(1.0e7)**		0.03
2e-3	w8a	49749	300			**3268.55(7.9e5)**				
1e-1	svmguide3	1243	22	0.47(5.1e4)	**0.15(3.7e4)**	2.84(4.2e7)	1.96(6.3e6)			**13.39**
1e-2	svmguide3	1243	22	7.95(1.4e5)	**2.41(6.6e4)**		74.73(6.3e6)			**30.98**
1e-3	svmguide3	1243	22	122.45(1.0e6)	**33.68(1.0e6)**		2946.23(3.4e7)			**87.49**
1e-4	svmguide3	1243	22	2418.83(1.0e7)	**804.17(2.1e7)**					
1e-1	phishing	11055	68	1.59(1.6e4)	1.17(1.6e4)	2.36(4.6e5)	102.48(1.4e7)	**1.17(4.7e6)**		0.99
1e-2	phishing	11055	68	17.53(2.9e4)	**12.18(3.5e4)**	158.49(1.5e7)	2506.03(2.2e7)			**13.01**
1e-3	phishing	11055	68	216.38(3.9e5)	**154.79(3.9e5)**					
1e-4	phishing	11055	68	5049.56(1.0e7)	**3558.37(1.0e7)**					
1e-1	ijcnn1	49990	22	2.32(8.2e3)	**0.96(9.2e3)**	10.88(6.6e5)	103.03(8.7e6)	1.58(3.7e6)	368.77(4.2e7)	**1.64**
1e-2	ijcnn1	49990	22	24.26(2.5e4)	**10.00(1.3e4)**	728.57(2.7e7)	2315.87(4.8e6)			**72.85**
1e-3	ijcnn1	49990	22	298.45(1.6e5)	**179.40(1.3e6)**					
1e-4	ijcnn1	49990	22		**2750.09(5.2e6)**					
5e-2	covtype	581012	54	156.41(4.5e6)	**110.28(5.8e6)**	2152.50(4.2e7)	2894.39(4.5e6)			**19.52**
1e-2	covtype	581012	54	785.43(4.7e6)	**572.97(5.8e6)**					
2e-3	covtype	581012	54	4292.90(5.8e6)	**3142.24(5.8e6)**					