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

October 28, 2023

A $\textit{Rule}-\textit{DBD}\sqrt[4]{K}$ for ERM with Non-convex Loss Functions

In a similar spirit as $\textit{Rule}-\textit{DBD}\sqrt{k}$, we also present the deterministic rule $\textit{Rule}-\textit{DBD}\sqrt[4]{K}$ which achieves nearly identical computational guarantees (to within a constant factor) as $\textit{Rule}-\textit{SBD}\sqrt[4]{K}$. First of all, let us recall $\textit{Rule}-\textit{DBD}\sqrt[4]{K}$ in Definition 4.5.

Definition A.1. $\textit{Rule}-\textit{DBD}\sqrt[4]{K}$. 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 $\textit{Rule}-\textit{DBD}\sqrt[4]{K}$ 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 $\textit{Rule}-\textit{DBD}\sqrt[4]{K}$ the Taylor points are updated less often as K grows (in a different way but with similar effect as in $\textit{Rule}-\textit{SBD}\sqrt[4]{K}$). Similar to the case of $\textit{Rule}-\textit{SBD}\sqrt[4]{K}$, we have:

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

Proof. The proof it is nearly identical to that of Proposition 4.2. For the initial iteration of Algorithm 2.1 the number of flops is $O(\mathrm{fLMO} + np^2)$. After the initial iteration, the number of flops in the first K iteration is

$$O\left(K \cdot \mathrm{fLMO} + np^2 + \sum_{i=1}^{K} (\beta_i + 1)p^2\right) \leq O(K \cdot \left(\mathrm{fLMO} + p^2\right) + K^{3/4} \cdot np^2\Big) \ .$$

Here the left-handsider follows from Proposition 2.1 and the inequality follows due to Proposition A.1. □

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

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

Corollary A.1. *Let*

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

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

$$O\left((\text{fLMO} \; + \; p^2)\left(\frac{(F(x^0) - F(x^*)) + \hat{L}D^3 + LD^2}{\epsilon}\right)^2 \; + \; p^2\left(\frac{(F(x^0) - F(x^*)) + \hat{L}D^3 + LD^2}{\epsilon}\right)^{3/2} \right) \; .$$

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 with non-convex losses. In the table $\epsilon_0 := F(x^0) - F(x^\star)$, $c_1 := LD^2$, and $c_2 := \hat{L}D^3$.

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.1)	$\mathbb{E}_{\hat{k}\sim\mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$	$O\left((\text{fLMO} + p^2) \cdot \frac{(\epsilon_0 + c_1 + c_2)^2}{\epsilon^2} + np^2 \cdot \frac{(\epsilon_0 + c_1 + c_2)^{3/2}}{\epsilon^{3/2}}\right)$
Standard Frank-Wolfe	$\mathbb{E}_{\hat{k}\sim\mathcal{U}([K])}[\mathcal{G}(x^{\hat{k}})] \leq \epsilon$	$O\left((\text{fLMO} + np) \cdot \frac{(\epsilon_0 + c_1)^2}{\epsilon^2}\right)$

Now 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:

$$\sum_{k=0}^{K} \frac{\mathcal{G}(x^k)}{K+1} \leq \frac{LD^2 + 2\varepsilon_0}{2\sqrt{K+1}} + \frac{1}{K+1}\sum_{k=0}^{K}(\nabla F(x^k) - g^k)^\top (s^k - \bar{s}^k) \; . \tag{A-2}$$

Applying Lemma 3.12 to (A-2), we obtain:

$$\sum_{k=0}^{K} \frac{\mathcal{G}(x^k)}{K+1} \leq \frac{2\epsilon_0 + LD^2}{2\sqrt{K+1}} + \frac{\hat{L}D^3}{2n(K+1)}\sum_{k=0}^{K}\sum_{i=1}^{n}\Big(\sum_{j=\tau_i^k}^{k-1}\gamma_j\Big)^2. \tag{A-3}$$

Note that $\gamma_j = 1/\sqrt{K+1}$ for any j, then

$$\sum_{k=0}^{K} \frac{\mathcal{G}(x^k)}{K+1} \leq \frac{2\epsilon_0 + LD^2}{2\sqrt{K+1}} + \frac{\hat{L}D^3}{2n(K+1)^2}\sum_{k=0}^{K}\sum_{i=1}^{n}(k - \tau_i^k)^2. \tag{A-4}$$

Notice in *Rule−DBD* $\sqrt[4]{K}$, $k - \tau_i^k \leq \lfloor K^{1/4} \rfloor - 1$, then $(k - \tau_i^k)^2 \leq \sqrt{K+1}$. Therefore,

$$\sum_{k=0}^{K} \frac{\mathcal{G}(x^k)}{K+1} \leq \frac{2\epsilon_0 + LD^2}{2\sqrt{K+1}} + \frac{\hat{L}D^3}{2\sqrt{K+1}}. \tag{A-5}$$

This is exactly the second inequality in (A-1). □

B Adaptive Step-size

In this section we are going to introduce the adaptive-step size proposed in (6.1) and prove the worst-case convergence rates in the case of using the TUFW with *Rule*$-$*SBD*$\sqrt{k}$ on (1.6) with convex objectives. Other rules are similar and less complicated.

We first recall the adaptive step-size as follows:

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

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 $[0, \gamma_k]$. Define $x(\lambda) := x^k + \gamma(s^k - x^k)$ and then

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

It could be further proven that when $\gamma \in [0, \gamma_k]$, the first three terms of the right-hand side dominates. Therefore, the $\tilde{\gamma}_k$ defined in (6.1), 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 $\arg\min_{\gamma\in[0,\gamma_k]} F(x(\gamma))$. Due to this reason, the adaptive step-size yields more decrease of the objective value than the standard step-size. Actually we have the following theorem on the convergence rate of Algorithm 2.1 with the adaptive step-sizes.

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 (1.1) with adaptive step-sizes defined by (B-6) for all $k \geq 0$. Then for all $k \geq 1$ we have:*

$$\mathbb{E}[F(x^k) - F(x^\star)] \leq \frac{2LD^2 + 544\hat{L}D^3}{k+1} \ . \tag{B-8}$$

Proof of Theorem B.1. First of all, suppose that $x^1, x^2, \dots$ denote the iterates of the TUFW with adaptive step-sizes and $s^1, s^2, \dots$ denote the outputs of the linear minimization oracle on $x^1, x^2, \dots$. We define $\delta_k := F(x^{k+1}) - F(x^k + \gamma_k(s^k - x^k))$, the difference of using adaptive step-sizes and standard step-sizes. Similar with the proof of Lemma 3.13, we have

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

where the inequality is due to (3.13) in Lemma 3.13. Subtracting $F^\star$ from both sides of the above inequality chain, we arrive at:

$$\varepsilon_{k+1} \leq (1-\gamma_k)\varepsilon_k + \gamma_k\left(\nabla F(x^k) - g^k\right)^\top (s^k - x^\star) + \gamma_k^2 LD^2/2 + \delta_k \ ,$$

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

$$(k+1)(k+2)\varepsilon_{k+1} \leq 2(k+1)LD^2 + \sum_{t=1}^k 2(t+1)(\nabla F(x^t) - g^t)^\top(s^t - x^\star) + \sum_{t=0}^k (t+1)(t+2)\delta_t \ . \tag{B-9}$$

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

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

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). Since Assumption 1.1 holds, an upper bound of $|C_i(\alpha)|$ is as follows:

$$
\begin{aligned}
|C_i(\alpha)| &\leq \big\|\nabla f_i(x^k + \alpha(s^k - x^k)) - \nabla f_i(b_i) - \nabla^2 f_i(b_i)\big(x^k + \alpha(s^k - x^k) - b_i\big)\big\|_* \cdot \|s^k - x^k\| \\
&\leq \frac{\hat{L}}{2} \cdot \|x^k + \alpha(s^k - x^k) - b_i\|^2 \cdot \|s^k - x^k\| \\
&\leq \hat{L} \cdot \Big\|x^k - x^{\tau_i^k}\Big\|^2 \cdot \|s^k - x^k\| + \hat{L}\alpha^2 \cdot \|s^k - x^k\|^3 \\
&\leq \hat{L}D \cdot \Big\|x^k - x^{\tau_i^k}\Big\|^2 + \alpha^2 \hat{L}D^3
\end{aligned}
$$

Here the first inequality is due to Proposition 3.11.

Now, for any $k \geq 0$

$$
\begin{aligned}
\delta_k &\leq \frac{1}{n}\sum_{i=1}^n \int_{\alpha=\gamma_k}^{\tilde{\gamma}_k} C_i(\alpha)\mathrm{d}\alpha \leq \frac{\gamma_k}{n} \cdot \sum_{i=1}^n \max_{\alpha\in[0,\gamma_k]} |C_i(\alpha)| \\
&\leq \frac{\hat{L}D\gamma_k}{n} \sum_{i=1}^n \Big\|\tilde{x}^k - \tilde{x}^{\tau_i^k}\Big\|^2 + \gamma_k^3 \hat{L}D^3 \ .
\end{aligned}
$$

Furthermore we have

$$
\|x^k - x^{\tau_i^k}\|^2 \leq \Bigg(\sum_{j=\tau_i^k}^{k-1} \|x^{j+1} - x^j\|\Bigg)^2 \leq \Bigg(\sum_{j=\tau_i^k}^{k-1} \tilde{\gamma}_j D\Bigg)^2 \leq \Bigg(\sum_{j=\tau_i^k}^{k-1} \gamma_j D\Bigg)^2 ,
$$

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

$$
\delta_k \leq \frac{\gamma_k \hat{L}D^3}{n} \sum_{i=1}^n \big(\sum_{j=\tau_i^k}^{k-1} \gamma_j\big)^2 + \gamma_k^3 \hat{L}D^3 \ . \tag{B-12}
$$

Next, we provide an upper bound of $(\nabla F(x^k) - g^k)^\top (s^k - x^\star)$ for any $k \geq 1$. For any $k \geq 1$, $\big(\nabla F(x^k) - g^k\big)^\top (s^k - x^\star)$ can be rewritten as

$$
\frac{1}{n}\sum_{i=1}^n \Big(\nabla f_i(x^k) - \nabla f_i(x^{\tau_i^k}) - \nabla^2 f_i(x^{\tau_i^k})(x^k - x^{\tau_i^k})\Big)^\top (s^k - x^\star), \tag{B-13}
$$

which, since $\|s^k - x^\star\| \le D$, is smaller than or equal to

$$\frac{D}{n}\sum_{i=1}^{n}\left\|\nabla f_i(x^k) - \nabla f_i(x^{\tau_i^k}) - \nabla^2 f_i(x^{\tau_i^k})(x^k - x^{\tau_i^k})\right\|_* \le \frac{\hat{L}D}{2n}\sum_{i=1}^{n}\|x^k - x^{\tau_i^k}\|^2, \tag{B-14}$$

where the inequality is due to (3.6) in Proposition 3.11. Additionally, we have

$$\|x^k - x^{\tau_i^k}\|^2 \le \Big(\sum_{j=\tau_i^k}^{k-1}\|x^{j+1} - x^j\|\Big)^2 \le \Big(\sum_{j=\tau_i^k}^{k-1}\tilde{\gamma}_j D\Big)^2 \le \Big(\sum_{j=\tau_i^k}^{k-1}\gamma_j D\Big)^2 ,$$

and substituting this last bound into (B-14) yields

$$\left(\nabla F(x^k) - g^k\right)^\top (s^k - x^\star) \le \frac{\hat{L}D^3}{2n}\sum_{i=1}^{n}\Big(\sum_{j=\tau_i^k}^{k-1}\gamma_j\Big)^2 . \tag{B-15}$$

Substituting (B-12), (B-15), and $\gamma_t := \frac{2}{t+2}$ into (B-9) yields

$$(k+1)(k+2)\varepsilon_{k+1} \le 2(k+1)LD^2 + \sum_{t=1}^{k}4(t+1)\cdot\frac{\hat{L}D^3}{2n}\cdot\sum_{i=1}^{n}\Big(\sum_{j=\tau_i^t}^{t-1}\gamma_j\Big)^2 + \sum_{t=1}^{k}\frac{8\hat{L}D^3}{t+2} . \tag{B-16}$$

Using Lemma 3.14 we obtain that

$$\mathbb{E}\left[\sum_{i=1}^{n}\Big(\sum_{j=\tau_i^t}^{t-1}\gamma_j\Big)^2\right] \le \frac{134n}{t+2} .$$

With this inequalty, applying expectation on both sides of (B-16) yields

$$\begin{aligned}(k+1)(k+2)\mathbb{E}\varepsilon_{k+1} \le&\ 2(k+1)LD^2 + \sum_{t=1}^{k}536\hat{L}D^3 + 8k\hat{L}D^3\\ \le&\ 2(k+1)LD^2 + 544k\hat{L}D^3 .\end{aligned}$$

Now this inequality above can directly lead to (B-8).

□

C 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 (7.1) by inflating the value of λ to $\lambda' = 100\lambda$ (where recall λ was determined by cross validation). Table C-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 C-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 C-2: Comparison of average CPU runtimes (in seconds) required to achieve $\mathcal{G}(x^k) \leq \epsilon$ for methods on the logistic regression problem (7.1) 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 C-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 (7.2). 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)**					