Supplementary Material To ICML 2019 Submission
Learning and Data Selection in Big Datasets

A. Proofs

A.1. Proposition 1

Observe for $\boldsymbol{z}^{(k+1)}$ of (9) that $g_2(\boldsymbol{z}^{(k+1)}) = K$, which satisfies the constraint of optimization problem $(P2a)$. For index sequence j, introduced in Proposition 1, define $c_{j_i}^{(k)} := \ell\left(\boldsymbol{x}_{j_i}, f(\boldsymbol{x}_{j_i}), h^{(k)}(\boldsymbol{x}_{j_i})\right)$. By definition, $c_{j_1}^{(k)} < c_{j_2}^{(k)} < \cdots < c_{j_N}^{(k)}$. We use the following lemmas:

Lemma 1. *For any $m \leq K$, the solution of $(P2a)$ satisfies $z_{j_m}^{(k+1)} = 1$.*

Lemma 2. *For any $m > K$, the solution of $(P2a)$ satisfies $z_{j_m}^{(k+1)} = 0$.*

From Lemmas 1 and 2, $\|\boldsymbol{z}^{(k+1)}\|_0 = K$ and

$$z_i^{(k+1)} = \begin{cases} 1 & i = j_1, j_2, \ldots, j_K \\ 0 & i = j_{K+1}, \ldots, j_N \end{cases},$$

which completes the proof.

A.2. Proposition 2

Note that the constraint on $\boldsymbol{z}$ must be closed and convex, as a sufficient condition for convergence of BCD. Clearly this is not the case with $\boldsymbol{z} \in \{0,1\}^N$ in (Q_1). Leveraging the equivalence between $(P2)$ and its linear program relaxation, $(P3)$, the constraint $\boldsymbol{z} \in [0\ ,\ 1]^N$ is closed and convex. Since a unique minimizer is found at each update, convergence to a stationary point follows from the standard convergence results Bertsekas (1999)[Chap 2.7].

A.3. Proposition 4

We start by proving Proposition 4 for $d = 1$. To this end, we first introduce a variant of $(P2)$ in which we define K equidistant marks in $x \in \mathcal{X} = [0, T]$ and project $\mathcal{E}^\star$ to this set of marks, namely we replace every entry in $\mathcal{E}^\star$ by its closest mark (measured by the Euclidian distance). Moreover, we limit $\mathcal{H}$ to the class of L-Lipschitz functions passing through those marks. We first observe that the approximation error of the solution of $(P2)$ is upper bounded by that of the variant. In the following, we derive the bound of Proposition 4 using the variant problem.

Divide entire domain $\mathcal{X}$ by K marks to some $K-1$ disjoint sets $\{\mathcal{S}_i \mid \bigcup_{i\in[K-1]} \mathcal{S}_i = \mathcal{X}, \mathcal{S}_i \bigcap \mathcal{S}_j = \phi, \forall i, j \in [K-1]\}$.

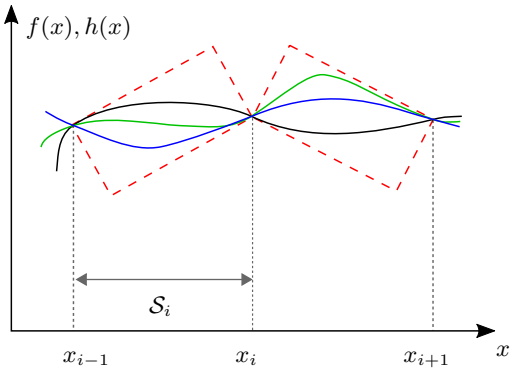

Figure A.1: Illustration of the functional class $\mathcal{F}$. Input space $\mathcal{X}$ is divided into disjoint sets $\{\mathcal{S}_i\}$. $\mathcal{F}$ is the set of all L-Lipschitz functions passing through samples/marks $\{x_i\}_{i\in[K]}$. All functions $f, h \in \mathcal{F}$ lie in the dashed red parallelograms. The slopes of these parallelograms are $\pm L$. Three possible functions are shown in the figure.

Define without loss of generality $\mathcal{S}_i = [x_{i-1}, x_i)$ for sorted x_i, and define $x_0 := 0$ and $x_{K-1} := T$. Note that $\mathcal{F}$ is the set of L-Lipschitz functions, samples are noiseless, and $\{x_i\}$ are in the compressed dataset. Figure A.1 illustrates the function class $\mathcal{F}$ and three potential examples for $f(x)$ and $h(x)$.

Define $\ell_x := \ell\left(x, f(x), h^\star(x)\right)$. Let μ_x be the probability measure on $\mathcal{X}$ that generates input samples x. We have

$$\mathbb{E}_x \ell_x = \int_{\mathcal{X}} \ell_x \, \mathrm{d}\mu_x = \sum_{i\in[K-1]} \int_{\mathcal{S}_i} \ell_{s_i} \, \mathrm{d}\mu_{s_i}\ , \qquad \text{(A.1)}$$

where $s_i \in \mathcal{S}_i$, $\{\mu_{s_i}\}$ are sub-probability measures on sets $\{\mathcal{S}_i\}$, and $\sum_{i\in[K-1]} \mu_{s_i} = 1$. From the extreme value theorem, there exists $\ell_{s_i}^{\max}$ for every interval $\mathcal{S}_i$ such that $\ell_{s_i} \leq \ell_{s_i}^{\max}, \forall s_i \in \mathcal{S}_i$. Therefore, $\mathbb{E}_x \ell_x \leq \max_i \ell_{s_i}^{\max}$. Consider the following lemma:

Lemma 3. *For our variant problem, $|f(\boldsymbol{x}) - h(\boldsymbol{x})| \leq 2L\|\boldsymbol{x}\|$ for all $\boldsymbol{x} \in \mathcal{S}_i$ and all i, where $\|\boldsymbol{x}\|$ is the L^2-norm of vector $\boldsymbol{x}$.*

The proof os Lemma 3 is straightforward after noting that $f(\boldsymbol{x}) - h(\boldsymbol{x})$ is a $2L$-Lipschitz function.

Consider loss function $\ell_x = |f(x) - h(x)|^2$. When $d = 1$, it is easy to see from Lemma 3 and figure A.1 that $\ell_{s_i}^{\max} \leq 4L^2(x_i - x_{i-1})^2$ for every set $\mathcal{S}_i$, where $x_i - x_{i-1}$ is the measure of set $\mathcal{S}_i$. Now, since sets $\{\mathcal{S}_i\}, i \in [K-1]$ have

the same measure (defined based on equidistant grid points), we have $x_i - x_{i-1} = T/(K-1)$, so

$$\mathbb{E}_x \ell_x \leq \max_i \ell_{s_i}^{\max} \leq \frac{4L^2T^2}{(K-1)^2}\,.$$

By setting $g(h^\star, \boldsymbol{z}^\star) \leq \mathbb{E}_x \ell_x \leq \delta$, we get $K \geq 1 + 2LT/\sqrt{\delta}$.

For $d > 1$, we can define equidistant marks on every coordinate of $\mathcal{X}$ and define a grid of $(K^{1/d}-1)^d$ disjoint sets $\{\mathcal{S}_i\}_i$, where we have assumed that $K^{1/d}$ is an integer number to avoid unnecessary notation complications. The distance between two consecutive marks on every coordinate is $T/(K^{1/d}-1)$, and therefore from Lemma 3

$$\mathbb{E}_x \leq \max_i \ell_{s_i}^{\max} \leq \left(2L\frac{T\sqrt{d}}{K^{1/d}-1}\right)^2\,.$$

By setting $g(h^\star, \boldsymbol{z}^\star) \leq \mathbb{E}_x \ell_x \leq \delta$, we get $K \geq \left(1 + 2LT\sqrt{d/\delta}\right)^d$. This completes the proof.

A.4. Lemma 1

Assume $\sum_{i\in[N]} z_i^{(k+1)} = \sum_{i\in[N]} z_{j_i}^{(k+1)} = M \geq K$. For $k=1$, if $z_{j_1}^{(k+1)} = 1$ the statement holds. If $z_{j_1}^{(k+1)} = 0$, then take any n for which $z_{j_n}^{(k+1)} = 1$ and observe that the following inequality holds by definition of index set j:

$$\frac{\sum_{i\in[N]} z_{j_i}^{(k+1)} c_{j_i}^{(k)}}{M} = \frac{\sum_{i\in[N]\setminus\{n\}} z_{j_i}^{(k+1)} c_{j_i}^{(k)} + c_{j_n}^{(k)}}{M} \geq \frac{\sum_{i\in[N]\setminus\{n\}} z_{j_i}^{(k+1)} c_{j_i}^{(k)} + c_{j_1}^{(k)}}{M}\,,$$

since $c_{j_1}^{(k)} < c_{j_n}^{(k)}$ for any n. This completes the proof for $k=1$. For $k = 2 \leq K$, if $z_{j_2}^{(k+1)} = 1$ the statement holds. If $z_{j_2}^{(k+1)} = 0$, then take any $n \geq 3$ for which $z_{j_n}^{(k+1)} = 1$. Use $z_{j_1}^{(k+1)} = 1$ and observe that

$$\frac{\sum\limits_{i\in[N]} z_{j_i}^{(k+1)} c_{j_i}^{(k)}}{M} = \frac{c_{j_1}^{(k)} + c_{j_n}^{(k)} + \sum\limits_{i\in[N]\setminus\{1,n\}} z_{j_i}^{(k+1)} c_{j_i}^{(k)}}{M} \geq \frac{c_{j_1}^{(k)} + c_{j_2}^{(k)} + \sum\limits_{i\in[N]\setminus\{1,n\}} z_{j_i}^{(k+1)} c_{j_i}^{(k)}}{M}$$

since $c_{j_2}^{(k)} < c_{j_n}^{(k)}$ for any $n > 2$. We can use the same arguments recursively to prove that $z_{j_m}^{(k+1)} = 1$ for any $m \leq K$.

A.5. Lemma 2

Assume $\sum_{i\in[N]} z_{j_i}^{(k+1)} = M \geq K$. By Lemma 1, $z_{j_i}^{(k+1)} = 1$ for all $i \leq K$. We should show that

$$\frac{\sum_{i\in[K]} c_{j_i}^{(k)}}{K} \leq \frac{\sum_{i\in[K]} c_{j_i}^{(k)} + \sum_{i=K+1}^{N} z_{j_i}^{(k+1)} c_{j_i}^{(k)}}{M}$$

for any $z_{j_i}^{(k+1)}$. This is clearly true as the left-hand-side is the average of the K smallest values of the loss function on dataset of size N. In particular,

$$\begin{aligned}
&\frac{\sum_{i\in[K]} c_{j_i}^{(k)} + \sum_{i=K+1}^{N} z_{j_i}^{(k+1)} c_{j_i}^{(k)}}{M} \\
&\geq \frac{\sum_{i\in[K]} c_{j_i}^{(k)} + \overbrace{c_{j_K}^{(k)} + c_{j_K}^{(k)} + \ldots + c_{j_K}^{(k)}}^{M-K}}{M} \\
&= \frac{\sum_{i\in[K]} c_{j_i}^{(k)}}{K} \\
&\qquad + \frac{\overbrace{(M-K)\,K c_{j_K}^{(k)} - (M-K)\sum_{i\in[K]} c_{j_i}^{(k)}}^{\geq 0}}{MK} \\
&\overset{(a)}{\geq} \frac{\sum_{i\in[K]} c_{j_i}^{(k)}}{K}\,,
\end{aligned}$$

where (a) holds as $K c_{j_K}^{(k)} \geq \sum_{i\in[K]} c_{j_i}^{(k)}$. This completes the proof.

A.6. Optimality of (8)

To prove the optimality of (8), recall that $\mathbf{1}^T\boldsymbol{z} = K$ in $(P2a)$ is of the form $\boldsymbol{A}\boldsymbol{z} = B$, where $\boldsymbol{A}$ is a totally unimodular matrix, and B is an integer. Thus, optimization problem(8) is equivalent to $(P2a)$, and the linear program relaxation is optimal.

B. Additional Examples

The following example shows the generality of Assumptions 1 and 2.

Example 4. *Let $\mathcal{P}$ denote the space of polynomial functions on $\mathbb{R}$, $f(x) = e^x (\in \mathcal{P})$, $h(x) = \sum_{n=0}^{N-1} x^n/n!$ $(\in \mathcal{P})$ be the first $N (< \infty)$ terms of the Taylor expansion of $f(x)$, and $\ell_n(h) := |f(x_n) - h(x_n)|^2$. $\ell_n(h)$ is compatible with Assumption 1. Moreover, for almost any $x_n, x_m \in \mathbb{R}$ (except a set of Lebesgue measure 0) such that $x_n \neq x_m$, we have $\ell_n(h) \neq \ell_m(h)$, so Assumption 2 holds.*

This example be easily generalized to the class of problems we study in this paper.

References

D.P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 2 edition, 1999.