

---

# Supplemental Materials for Faithful Group Shapley Value

---

**Kiljae Lee\***  
The Ohio State University  
lee.10428@osu.edu

**Ziqi Liu\***  
Carnegie Mellon University  
ziqiliu2@andrew.cmu.edu

**Weijing Tang**  
Carnegie Mellon University  
weijingt@andrew.cmu.edu

**Yuan Zhang<sup>†</sup>**  
The Ohio State University  
yzhanghf@stat.osu.edu

## A Proofs

### Contents

A.1	Proof of Proposition 1 . . . . .	1
A.1.1	Technical lemmas used in the proof of Proposition 1 . . . . .	4
A.2	Proof of Theorem 1 . . . . .	7
A.3	Proof of Lemma 1 . . . . .	8
A.4	Proof of Theorem 2 . . . . .	9
A.4.1	Technical lemmas used in the proof of Theorem 2 . . . . .	12
A.5	Proof of Theorem 3 . . . . .	16
A.5.1	Technical lemmas used in the proof of Theorem 3 . . . . .	17
A.6	Proof of Proposition 2 . . . . .	19
A.6.1	Preliminary first-order stability results and proof of Lemma 11 . . . . .	21
A.6.2	Technical lemmas used in the proof of Proposition 2 . . . . .	26
A.7	Proof of Proposition 3 (algorithmic stability of Influence Function (IF)) . . . . .	31
A.7.1	Lemmas used in the proof of Proposition 3 . . . . .	32

### A.1 Proof of Proposition 1

We first introduce some notation and preliminary results. Proposition 1 considers a group  $S_k$  (from an initial set of  $K + 1$  groups  $S_0, \dots, S_K$ ) which is split into two non-empty, disjoint subgroups,  $S'_k$  and  $S''_k$ . We denote their cardinalities as  $s_l = |S_l|$  (for  $l = 0, \dots, K$ ),  $s'_k = |S'_k|$ , and  $s''_k = |S''_k|$ , such that  $s_k = s'_k + s''_k$ . Let  $\bar{U}(s)$  be the expected utility function defined in Proposition 1, which depends only on the dataset size  $s$ . We define an auxiliary function  $\Delta(x)$  for any  $x \in \mathbb{N}$  (where  $x$  will typically represent the sum of cardinalities of other groups in a coalition):

$$\Delta(x) := \bar{U}(x + s'_k) + \bar{U}(x + s''_k) - \bar{U}(x) - \bar{U}(x + s'_k + s''_k).$$

The proof of Proposition 1 will rely on the following technical lemmas. Their proofs, which build upon the prudence condition (3), are deferred to Appendix A.1.1.

---

\*Lee and Liu equally contributed; they are co-first authors and were listed alphabetically.

<sup>†</sup>Corresponding author.

**Lemma 2.** [Prudence Implication for First Differences] Suppose the expected utility function  $\bar{U}(s)$  satisfies the prudence condition (3). Let  $\Delta\bar{U}(t) := \bar{U}(t+1) - \bar{U}(t)$  for  $t \geq 0$ . Then, for any integer  $x \geq 0$  and positive integers  $a, b \in \mathbb{N}_+$ , it holds that

$$\Delta\bar{U}(x) + \Delta\bar{U}(x+a+b) > \Delta\bar{U}(x+a) + \Delta\bar{U}(x+b).$$

(This property indicates that the first-order difference  $\Delta\bar{U}(x)$  is a strictly convex function.)

**Lemma 3.** [Strict Monotonicity of  $\Delta(x)$ ] If  $\bar{U}(s)$  satisfies the prudence condition (3) (which implies Lemma 2), then  $\Delta(x)$  is a strictly decreasing function of  $x$ . That is, for any  $x_2 > x_1 \geq 0$ , we have  $\Delta(x_1) > \Delta(x_2)$ .

**Lemma 4.** [Sum Symmetrization Identity] Let  $[n] := \{1, \dots, n\}$  be the set of player indices, and let  $f : 2^{[n]} \rightarrow \mathbb{R}$  be a function that assigns a real value to each subset of  $[n]$ . Then, for any integer  $m$  such that  $0 \leq m \leq \lfloor n/2 \rfloor$ , the following identity holds:

$$\sum_{\substack{I \subseteq [n] \\ |I|=m}} f(I) + \sum_{\substack{I \subseteq [n] \\ |I|=n-m}} f(I) = \frac{1}{\binom{n-m}{m}} \sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} [f(I) + f(I \cup J)].$$

We now proceed to prove Proposition 1.

*Proof of Proposition 1.* Let  $N_0 = \{0, \dots, K\}$  be the set of indices for the initial  $K+1$  groups  $S_0, \dots, S_K$ . The Group Shapley Value for  $S_k$  can be re-expressed in terms of  $\bar{U}$ , as follows:

$$\begin{aligned} \mathbb{E}[\text{GSV}(S_k)] &= \sum_{I^* \subseteq N_0 \setminus \{k\}} \frac{|I^*|!(K - |I^*|)!}{(K+1)!} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) \right] \\ &= \frac{1}{K+1} \sum_{m^*=0}^K \frac{1}{\binom{K}{m^*}} \sum_{\substack{I^* \subseteq N_0 \setminus \{k\} \\ |I^*|=m^*}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) \right]. \end{aligned} \quad (1)$$

After splitting  $S_k$  into  $S'_k$  and  $S''_k$ , we have  $K+2$  groups. Let  $N' = (N_0 \setminus \{k\}) \cup \{k', k''\}$  be the set of indices for these  $K+2$  groups, where  $k'$  and  $k''$  index  $S'_k$  and  $S''_k$  respectively. The GSVs for  $S'_k$  and  $S''_k$  are:

$$\begin{aligned} \mathbb{E}[\text{GSV}(S'_k)] &= \sum_{I \subseteq N' \setminus \{k'\}} \frac{|I|!(K+1 - |I|)!}{(K+2)!} \left[ \bar{U} \left( \sum_{l \in I} s_l + s'_k \right) - \bar{U} \left( \sum_{l \in I} s_l \right) \right] \\ &= \sum_{m^*=0}^{K+1} \frac{1}{K+2} \frac{1}{\binom{K+1}{m^*}} \sum_{\substack{I^* \subseteq N' \setminus \{k'\} \\ |I^*|=m^*}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s'_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) \right], \\ \mathbb{E}[\text{GSV}(S''_k)] &= \sum_{I \subseteq N' \setminus \{k''\}} \frac{|I|!(K+1 - |I|)!}{(K+2)!} \left[ \bar{U} \left( \sum_{l \in I} s_l + s''_k \right) - \bar{U} \left( \sum_{l \in I} s_l \right) \right] \\ &= \sum_{m^*=0}^{K+1} \frac{1}{K+2} \frac{1}{\binom{K+1}{m^*}} \sum_{\substack{I^* \subseteq N' \setminus \{k''\} \\ |I^*|=m^*}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) \right]. \end{aligned}$$

Let  $I_0 = N_0 \setminus \{k\}$  be the set of  $K$  "other" groups. To analyze  $\mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)]$ , we examine the terms associated with each  $I^* \subseteq I_0$ . Let  $m^* := |I^*|$ . For a given  $I^*$ , its coalition can be joined by  $S'_k$  alone,  $S''_k$  alone, or by  $S'_k$  and  $S''_k$  in either order. Collecting the marginal contributions for  $S'_k$  and  $S''_k$  across these scenarios for a fixed  $I^*$ , we obtain:

$$\begin{aligned} &\frac{1}{K+2} \left\{ \frac{1}{\binom{K+1}{m^*}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s'_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) + \bar{U} \left( \sum_{l \in I^*} s_l + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) \right] \right. \\ &\quad \left. + \frac{1}{\binom{K+1}{m^*+1}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s'_k + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l + s''_k \right) + \bar{U} \left( \sum_{l \in I^*} s_l + s'_k + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l + s'_k \right) \right] \right\}. \end{aligned}$$

Summing over all  $I^* \subseteq I_0$ , the total expected value is:

$$\begin{aligned} & \mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] \\ &= \sum_{I^* \subseteq I_0} \frac{1}{K+2} \left\{ \frac{1}{\binom{K+1}{|I^*|}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s'_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) + \bar{U} \left( \sum_{l \in I^*} s_l + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l \right) \right] \right. \\ & \quad \left. + \frac{1}{\binom{K+1}{|I^*|+1}} \left[ \bar{U} \left( \sum_{l \in I^*} s_l + s'_k + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l + s''_k \right) + \bar{U} \left( \sum_{l \in I^*} s_l + s'_k + s''_k \right) - \bar{U} \left( \sum_{l \in I^*} s_l + s'_k \right) \right] \right\}. \end{aligned}$$

Let  $s_{I^*} = \sum_{l \in I^*} s_l$ . The above form can be simplified into:

$$\begin{aligned} & \mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] \\ &= \sum_{I^* \subseteq I_0} \frac{1}{K+2} \left\{ \frac{1}{\binom{K+1}{|I^*|}} [\bar{U}(s_{I^*} + s'_k) + \bar{U}(s_{I^*} + s''_k) - 2\bar{U}(s_{I^*})] \right. \\ & \quad \left. + \frac{1}{\binom{K+1}{|I^*|+1}} [2\bar{U}(s_{I^*} + s_k) - \bar{U}(s_{I^*} + s''_k) - \bar{U}(s_{I^*} + s'_k)] \right\}. \end{aligned}$$

For  $\mathbb{E}[\text{GSV}(S_k)]$  in Eq. (1), using the identity  $\frac{1}{K+1} \frac{1}{\binom{K}{m^*}} = \frac{1}{K+2} \left( \frac{1}{\binom{K+1}{m^*}} + \frac{1}{\binom{K+1}{m^*+1}} \right)$ , we have:

$$\begin{aligned} \mathbb{E}[\text{GSV}(S_k)] &= \sum_{I^* \subseteq I_0} \frac{1}{K+1} \frac{1}{\binom{K}{|I^*|}} [\bar{U}(s_{I^*} + s_k) - \bar{U}(s_{I^*})] \\ &= \sum_{I^* \subseteq I_0} \frac{1}{K+2} \left( \frac{1}{\binom{K+1}{|I^*|}} + \frac{1}{\binom{K+1}{|I^*|+1}} \right) [\bar{U}(s_{I^*} + s_k) - \bar{U}(s_{I^*})]. \end{aligned}$$

Then, the difference in the expected GSVs is:

$$\begin{aligned} & \mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] - \mathbb{E}[\text{GSV}(S_k)] \\ &= \sum_{I^* \subseteq I_0} \frac{1}{K+2} \left\{ \frac{1}{\binom{K+1}{|I^*|}} [\bar{U}(s_{I^*} + s'_k) + \bar{U}(s_{I^*} + s''_k) - 2\bar{U}(s_{I^*}) - \bar{U}(s_{I^*} + s_k) + \bar{U}(s_{I^*})] \right. \\ & \quad \left. + \frac{1}{\binom{K+1}{|I^*|+1}} [2\bar{U}(s_{I^*} + s_k) - \bar{U}(s_{I^*} + s''_k) - \bar{U}(s_{I^*} + s'_k) - \bar{U}(s_{I^*} + s_k) + \bar{U}(s_{I^*})] \right\} \\ &= \sum_{I^* \subseteq I_0} \frac{1}{K+2} \left( \frac{1}{\binom{K+1}{m^*}} - \frac{1}{\binom{K+1}{m^*+1}} \right) \Delta(s_{I^*}). \end{aligned}$$

We further group the terms with the same cardinality  $m^* = |I^*|$ , which ranges from 0 to  $K$ . For each  $m^*$ , there are  $\binom{K}{m^*}$  such coalitions  $I^*$ . This yields:

$$\begin{aligned} & \mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] - \mathbb{E}[\text{GSV}(S_k)] \\ &= \frac{1}{K+2} \sum_{m^*=0}^K \binom{K}{m^*} \left( \frac{1}{\binom{K+1}{m^*}} - \frac{1}{\binom{K+1}{m^*+1}} \right) \frac{\sum_{I^*: |I^*|=m^*} \Delta(s_{I^*})}{\binom{K}{m^*}} \\ &= \frac{1}{(K+1)(K+2)} \sum_{m^*=0}^K (K-2m^*) \frac{\sum_{I^*: |I^*|=m^*} \Delta(s_{I^*})}{\binom{K}{m^*}}, \end{aligned} \tag{2}$$

where the last step uses the fact  $\binom{K}{m^*} \left( \frac{1}{\binom{K+1}{m^*}} - \frac{1}{\binom{K+1}{m^*+1}} \right) = \frac{K-2m^*}{K+1}$ .

To determine the sign of the RHS of (2), we define an auxiliary function

$$g(I^*) := \frac{K-2|I^*|}{\binom{K}{|I^*|}} \Delta(s_{I^*})$$

for all  $I^* \subseteq I_0$ . We have

$$\mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] - \mathbb{E}[\text{GSV}(S_k)] = \frac{1}{(K+1)(K+2)} \sum_{I^* \subseteq I_0} g(I^*). \quad (3)$$

We can rewrite the sum  $\sum_{I^* \subseteq I_0} g(I^*)$  by pairing up terms corresponding to coalitions of sizes  $m^*$  and  $K - m^*$ , for each  $m^* = 0, \dots, \lfloor (K-1)/2 \rfloor$ . We have

$$\sum_{I^* \subseteq I_0} g(I^*) = \sum_{m^*=0}^{\lfloor (K-1)/2 \rfloor} \left( \sum_{\substack{I^* \subseteq I_0 \\ |I^*|=m^*}} g(I^*) + \sum_{\substack{J^* \subseteq I_0 \\ |J^*|=K-m^*}} g(J^*) \right).$$

(If  $K$  is even, the middle term where  $m^* = K/2$  has  $K - 2m^* = 0$ , so  $g(I^*) = 0$  for  $|I^*| = K/2$ .) Applying Lemma 4 to each parenthesized pair of sums:

$$\sum_{\substack{I^* \subseteq I_0 \\ |I^*|=m^*}} g(I^*) + \sum_{\substack{J^* \subseteq I_0 \\ |J^*|=K-m^*}} g(J^*) = \frac{1}{\binom{K-m^*}{m^*}} \sum_{\substack{I^* \subseteq I_0 \\ |I^*|=m^*}} \sum_{\substack{L^* \subseteq I_0 \setminus I^* \\ |L^*|=K-2m^*}} [g(I^*) + g(I^* \cup L^*)].$$

Substituting  $g(I^*) = \frac{K-2m^*}{\binom{K}{m^*}} \Delta(s_{I^*})$  and noting that  $|I^* \cup L^*| = K - m^*$ , we have  $g(I^* \cup L^*) = \frac{K-2(K-m^*)}{\binom{K}{K-m^*}} \Delta(s_{I^* \cup L^*}) = -\frac{(K-2m^*)}{\binom{K}{m^*}} \Delta(s_{I^* \cup L^*})$ . Thus, the term  $[g(I^*) + g(I^* \cup L^*)]$  becomes  $\frac{K-2m^*}{\binom{K}{m^*}} [\Delta(s_{I^*}) - \Delta(s_{I^* \cup L^*})]$ . So,  $\mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] - \mathbb{E}[\text{GSV}(S_k)]$  is:

$$\frac{1}{(K+1)(K+2)} \sum_{m^*=0}^{\lfloor (K-1)/2 \rfloor} \frac{1}{\binom{K-m^*}{m^*}} \frac{K-2m^*}{\binom{K}{m^*}} \sum_{\substack{I^* \subseteq I_0 \\ |I^*|=m^*}} \sum_{\substack{L^* \subseteq I_0 \setminus I^* \\ |L^*|=K-2m^*}} [\Delta(s_{I^*}) - \Delta(s_{I^* \cup L^*})].$$

For  $m^* < K/2$ , we have:

1. The coefficient  $\frac{K-2m^*}{(K+1)(K+2)\binom{K-m^*}{m^*}\binom{K}{m^*}}$  is strictly positive.
2. The set  $L^*$  is non-empty (since  $K - 2m^* > 0$ ). Assuming at least some  $s_l > 0$  for  $l \in L^*$ , we have  $s_{I^* \cup L^*} > s_{I^*}$ .
3. By Lemma 3, we have  $\Delta(s_{I^*}) > \Delta(s_{I^* \cup L^*})$ .

Combining these observations shows  $\mathbb{E}[\text{GSV}(S'_k)] + \mathbb{E}[\text{GSV}(S''_k)] > \mathbb{E}[\text{GSV}(S_k)]$  and completes the proof of Proposition 1.  $\square$

### A.1.1 Technical lemmas used in the proof of Proposition 1

**Lemma 2.** [Prudence Implication for First Differences] Suppose the expected utility function  $\bar{U}(s)$  satisfies the prudence condition (3). Let  $\Delta \bar{U}(t) := \bar{U}(t+1) - \bar{U}(t)$  for  $t \geq 0$ . Then, for any integer  $x \geq 0$  and positive integers  $a, b \in \mathbb{N}_+$ , it holds that

$$\Delta \bar{U}(x) + \Delta \bar{U}(x+a+b) > \Delta \bar{U}(x+a) + \Delta \bar{U}(x+b).$$

(This property indicates that the first-order difference  $\Delta \bar{U}(x)$  is a strictly convex function.)

*Proof of Lemma 2.* Define the second-order difference as  $\Delta^2 \bar{U}(t) := \Delta \bar{U}(t+1) - \Delta \bar{U}(t)$ . The prudence condition on  $\bar{U}(s)$  is given by (3):  $\Delta^3 \bar{U}(x) := \bar{U}(x+3) - 3\bar{U}(x+2) + 3\bar{U}(x+1) - \bar{U}(x) > 0$  for any valid  $x$ . We can express  $\Delta^3 \bar{U}(x)$  as follows:

$$\begin{aligned} \Delta^3 \bar{U}(x) &= (\bar{U}(x+3) - \bar{U}(x+2)) - 2(\bar{U}(x+2) - \bar{U}(x+1)) + (\bar{U}(x+1) - \bar{U}(x)) \\ &= \Delta \bar{U}(x+2) - 2\Delta \bar{U}(x+1) + \Delta \bar{U}(x) \\ &= \Delta^2 \bar{U}(x+1) - \Delta^2 \bar{U}(x). \end{aligned}$$

Since  $\Delta^3 \bar{U}(x) > 0$ , we have  $\Delta^2 \bar{U}(x+1) > \Delta^2 \bar{U}(x)$ . As this holds for any valid  $x$ , it implies that the function  $\Delta^2 \bar{U}(t)$  is strictly increasing in  $t$ .

We want to prove that for any integer  $x \geq 0$  and positive integers  $a, b \in \mathbb{N}_+$ :

$$\Delta \bar{U}(x) + \Delta \bar{U}(x+a+b) > \Delta \bar{U}(x+a) + \Delta \bar{U}(x+b).$$

This inequality can be rearranged to:

$$\Delta \bar{U}(x+a+b) - \Delta \bar{U}(x+a) > \Delta \bar{U}(x+b) - \Delta \bar{U}(x). \quad (4)$$

The left-hand side (LHS) of (4) can be written as a sum of second-order differences:

$$\begin{aligned} \Delta \bar{U}(x+a+b) - \Delta \bar{U}(x+a) &= \sum_{j=0}^{b-1} (\Delta \bar{U}(x+a+j+1) - \Delta \bar{U}(x+a+j)) \\ &= \sum_{j=0}^{b-1} \Delta^2 \bar{U}(x+a+j). \end{aligned}$$

Similarly, the right-hand side (RHS) of (4) is:

$$\begin{aligned} \Delta \bar{U}(x+b) - \Delta \bar{U}(x) &= \sum_{j=0}^{b-1} (\Delta \bar{U}(x+j+1) - \Delta \bar{U}(x+j)) \\ &= \sum_{j=0}^{b-1} \Delta^2 \bar{U}(x+j). \end{aligned}$$

Since  $a \in \mathbb{N}_+$ , we have  $a \geq 1$ . Therefore, for each  $j \in \{0, \dots, b-1\}$ , the argument  $x+a+j$  is strictly greater than  $x+j$ . Because  $\Delta^2 \bar{U}(t)$  is a strictly increasing function of  $t$ , it follows that for each term in the summations:

$$\Delta^2 \bar{U}(x+a+j) > \Delta^2 \bar{U}(x+j).$$

Summing these strict inequalities from  $j = 0$  to  $b-1$  (since  $b \in \mathbb{N}_+$ , there is at least one term in the sum):

$$\sum_{j=0}^{b-1} \Delta^2 \bar{U}(x+a+j) > \sum_{j=0}^{b-1} \Delta^2 \bar{U}(x+j).$$

This directly implies that the LHS of (4) is strictly greater than its RHS:

$$\Delta \bar{U}(x+a+b) - \Delta \bar{U}(x+a) > \Delta \bar{U}(x+b) - \Delta \bar{U}(x).$$

Rearranging this yields the desired result:

$$\Delta \bar{U}(x) + \Delta \bar{U}(x+a+b) > \Delta \bar{U}(x+a) + \Delta \bar{U}(x+b).$$

This completes the proof.  $\square$

**Lemma 3.** [Strict Monotonicity of  $\Delta(x)$ ] If  $\bar{U}(s)$  satisfies the prudence condition (3) (which implies Lemma 2), then  $\Delta(x)$  is a strictly decreasing function of  $x$ . That is, for any  $x_2 > x_1 \geq 0$ , we have  $\Delta(x_1) > \Delta(x_2)$ .

*Proof of Lemma 3.* We want to show that  $\Delta(x)$  is a strictly decreasing function of  $x$ . This is equivalent to proving that  $\Delta(x) - \Delta(x+1) > 0$  for any  $x \geq 0$  (assuming  $x$  and  $x+1$  lead to valid arguments for  $\bar{U}$  as per the lemma statement).

Let the first-order difference of  $\bar{U}$  be defined as  $\Delta \bar{U}(t) := \bar{U}(t+1) - \bar{U}(t)$ , for  $t \geq 0$ , ensuring all arguments to  $\bar{U}$  are valid. Recall the definition of  $\Delta(x)$ :

$$\Delta(x) = \bar{U}(x+s'_k) + \bar{U}(x+s''_k) - \bar{U}(x) - \bar{U}(x+s'_k+s''_k).$$

Then, the difference  $\Delta(x) - \Delta(x+1)$  is:

$$\begin{aligned} \Delta(x) - \Delta(x+1) &= [\bar{U}(x+s'_k) + \bar{U}(x+s''_k) - \bar{U}(x) - \bar{U}(x+s'_k+s''_k)] \\ &\quad - [\bar{U}(x+1+s'_k) + \bar{U}(x+1+s''_k) - \bar{U}(x+1) - \bar{U}(x+1+s'_k+s''_k)]. \end{aligned}$$

We regroup the terms and leverage the definition of  $\Delta\bar{U}(t)$ , this becomes:

$$\Delta(x) - \Delta(x+1) = \Delta\bar{U}(x) + \Delta\bar{U}(x + s'_k + s''_k) - [\Delta\bar{U}(x + s'_k) + \Delta\bar{U}(x + s''_k)].$$

Since  $S'_k$  and  $S''_k$  are non-empty,  $s'_k$  and  $s''_k$  are positive integers (i.e.,  $a, b \in \mathbb{N}_+$ ). By Lemma 2, we have:

$$\Delta\bar{U}(x) + \Delta\bar{U}(x + s'_k + s''_k) > \Delta\bar{U}(x + s'_k) + \Delta\bar{U}(x + s''_k).$$

Therefore,

$$\Delta\bar{U}(x) + \Delta\bar{U}(x + s'_k + s''_k) - [\Delta\bar{U}(x + s'_k) + \Delta\bar{U}(x + s''_k)] > 0.$$

This implies  $\Delta(x) - \Delta(x+1) > 0$ , so  $\Delta(x) > \Delta(x+1)$ . Since this holds for any valid  $x \geq 0$ ,  $\Delta(x)$  is a strictly decreasing function of  $x$ . Consequently, for any  $x_2 > x_1 \geq 0$ , we have  $\Delta(x_1) > \Delta(x_2)$ .  $\square$

**Lemma 4.** [Sum Symmetrization Identity] Let  $[n] := \{1, \dots, n\}$  be the set of player indices, and let  $f : 2^{[n]} \rightarrow \mathbb{R}$  be a function that assigns a real value to each subset of  $[n]$ . Then, for any integer  $m$  such that  $0 \leq m \leq \lfloor n/2 \rfloor$ , the following identity holds:

$$\sum_{\substack{I \subseteq [n] \\ |I|=m}} f(I) + \sum_{\substack{I \subseteq [n] \\ |I|=n-m}} f(I) = \frac{1}{\binom{n-m}{m}} \sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} [f(I) + f(I \cup J)].$$

*Proof of Lemma 4.* We start by analyzing the right-hand side (RHS) of the proposed identity:

$$\text{RHS} = \frac{1}{\binom{n-m}{m}} \sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} [f(I) + f(I \cup J)].$$

We can split the sum inside the square brackets into two parts:

$$\text{RHS} = \frac{1}{\binom{n-m}{m}} \left[ \sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} f(I) + \sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} f(I \cup J) \right].$$

Let's analyze the first double summation:  $\sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} f(I)$ . For any fixed subset  $I$  such that  $|I| = m$ , the term  $f(I)$  is constant with respect to the inner sum over  $J$ . The number of ways to choose a subset  $J \subseteq [n] \setminus I$  with  $|J| = n - 2m$  is given by  $\binom{|[n] \setminus I|}{n-2m}$ . Since  $|[n] \setminus I| = n - m$ , this count is  $\binom{n-m}{n-2m}$ . Using the identity  $\binom{k}{r} = \binom{k}{k-r}$ , we have  $\binom{n-m}{n-2m} = \binom{n-m}{(n-m)-(n-2m)} = \binom{n-m}{m}$ . So, for each fixed  $I$  with  $|I| = m$ ,  $f(I)$  appears  $\binom{n-m}{m}$  times. Thus, the first double summation is:

$$\sum_{\substack{I \subseteq [n] \\ |I|=m}} \binom{n-m}{m} f(I).$$

Now, let's analyze the second double summation:  $\sum_{\substack{I \subseteq [n] \\ |I|=m}} \sum_{\substack{J \subseteq [n] \setminus I \\ |J|=n-2m}} f(I \cup J)$ . Let  $K = I \cup J$ .

Since  $I \cap J = \emptyset$ ,  $|I| = m$ , and  $|J| = n - 2m$ , it follows that  $|K| = m + (n - 2m) = n - m$ . We want to count how many times a specific subset  $K \subseteq [n]$  with  $|K| = n - m$  appears as  $I \cup J$  in this summation. For a fixed  $K$  (where  $|K| = n - m$ ), we need to find an  $I \subseteq K$  such that  $|I| = m$ . Once such an  $I$  is chosen,  $J$  is uniquely determined as  $J = K \setminus I$ . The size of this  $J$  will be  $|K \setminus I| = (n - m) - m = n - 2m$ . Also,  $J \subseteq [n] \setminus I$  is satisfied. The number of ways to choose such a subset  $I \subseteq K$  with  $|I| = m$  is  $\binom{|K|}{m} = \binom{n-m}{m}$ . So, for each fixed  $K$  with  $|K| = n - m$ , the term  $f(K)$  appears  $\binom{n-m}{m}$  times. Thus, the second double summation is:

$$\sum_{\substack{K \subseteq [n] \\ |K|=n-m}} \binom{n-m}{m} f(K).$$

(Renaming the dummy variable  $K$  to  $I$  for consistency with the LHS):

$$\sum_{\substack{I \subseteq [n] \\ |I|=n-m}} \binom{n-m}{m} f(I).$$

Substituting these back into the expression for the RHS:

$$\begin{aligned} \text{RHS} &= \frac{1}{\binom{n-m}{m}} \left[ \sum_{\substack{I \subseteq [n] \\ |I|=m}} \binom{n-m}{m} f(I) + \sum_{\substack{I \subseteq [n] \\ |I|=n-m}} \binom{n-m}{m} f(I) \right] \\ &= \sum_{\substack{I \subseteq [n] \\ |I|=m}} f(I) + \sum_{\substack{I \subseteq [n] \\ |I|=n-m}} f(I). \end{aligned}$$

This is exactly the left-hand side (LHS) of the identity stated in the lemma. The condition  $0 \leq m \leq \lfloor n/2 \rfloor$  ensures that  $n-2m \geq 0$  (so  $|J|$  is a valid size) and  $m \leq n-m$  (so  $\binom{n-m}{m}$  is well-defined and typically non-zero, unless  $n-m < m$  which is prevented by  $m \leq n/2$ . If  $m = n/2$ ,  $\binom{n-m}{m} = \binom{m}{m} = 1$ ). This completes the proof.  $\square$

## A.2 Proof of Theorem 1

*Proof of Theorem 1.* Let  $[n] = \{1, \dots, n\}$  be the set of indices for the data points. The utility function  $U$  is defined on subsets of  $[n]$ , and  $\Pi$  is a partition of  $[n]$ . The proof consists of two parts: sufficiency and uniqueness.

**Sufficiency.** We verify that the proposed valuation  $\nu_{U, \mathcal{D}, \Pi}(S) = \sum_{i \in S} \text{SV}(i)$  satisfies all five axioms from Definition 2. For any  $S_j \in \Pi$ :

1. **Null player:** If every subset  $S' \subseteq S_j$  satisfies  $U(S'' \cup S') = U(S'')$  for all  $S'' \subseteq [n] \setminus S'$ , then as a special case, for every element  $k \in S_j$ , it holds that  $U(S'' \cup \{k\}) = U(S'')$  for all  $S'' \subseteq [n] \setminus \{k\}$ . By the definition of individual Shapley value, we have  $\text{SV}(k) = 0$ . Thus,  $\nu_{U, \mathcal{D}, \Pi}(S_j) = \sum_{k \in S_j} \text{SV}(k) = 0$ .
2. **Symmetry:** Let  $S_1, S_2$  be the said sets and  $\sigma$  be the said mapping between them, all as described in the axiom. For every individual  $k \in S_1$ , we take  $S' = \{k\}$ . By the axiom, for all  $S'' \subseteq [n] \setminus \{k, \sigma(k)\}$ , it holds that  $U(S'' \cup \{k\}) = U(S'' \cup \{\sigma(k)\})$ . Then it is easy to verify that  $\text{SV}(k) = \text{SV}(\sigma(k))$ . Therefore,  $\nu_{U, \mathcal{D}, \Pi}(S_1) = \sum_{k \in S_1} \text{SV}(k) = \sum_{k \in S_1} \text{SV}(\sigma(k)) = \sum_{l \in S_2} \text{SV}(l) = \nu_{U, \mathcal{D}, \Pi}(S_2)$ .
3. **Linearity:** For utility functions  $U_1, U_2$  and scalars  $\alpha_1, \alpha_2$ , the individual Shapley value is linear:  $\text{SV}_{\alpha_1 U_1 + \alpha_2 U_2}(k) = \alpha_1 \text{SV}_{U_1}(k) + \alpha_2 \text{SV}_{U_2}(k)$ . Summing over  $k \in S_j$  preserves this linearity:  $\nu_{\alpha_1 U_1 + \alpha_2 U_2, \mathcal{D}, \Pi}(S_j) = \sum_{k \in S_j} \text{SV}_{\alpha_1 U_1 + \alpha_2 U_2}(k) = \alpha_1 \sum_{k \in S_j} \text{SV}_{U_1}(k) + \alpha_2 \sum_{k \in S_j} \text{SV}_{U_2}(k) = \alpha_1 \nu_{U_1, \mathcal{D}, \Pi}(S_j) + \alpha_2 \nu_{U_2, \mathcal{D}, \Pi}(S_j)$ .
4. **Efficiency:** For any partition  $\Pi = \{S_1, \dots, S_M\}$  of  $[n]$ , the axiom implies that  $\sum_{j=1}^M \nu_{U, \mathcal{D}, \Pi}(S_j) = \sum_{j=1}^M \sum_{k \in S_j} \text{SV}(k) = \sum_{k \in [n]} \text{SV}(k)$ . By the efficiency property of the individual Shapley value, we have  $\sum_{k \in [n]} \text{SV}(k) = U([n])$ .
5. **Faithfulness:** The proposed valuation  $\nu_{U, \mathcal{D}, \Pi}(S) = \sum_{i \in S} \text{SV}(i)$  depends only on the group  $S$  itself (and  $U, \mathcal{D}$ ), not on the specific partition  $\Pi$  that  $S$  belongs to. Thus, if  $S_0 \in \Pi_1 \cap \Pi_2$ , then  $\nu_{U, \mathcal{D}, \Pi_1}(S_0) = \sum_{i \in S_0} \text{SV}(i)$  and  $\nu_{U, \mathcal{D}, \Pi_2}(S_0) = \sum_{i \in S_0} \text{SV}(i)$ . These are equal, so the Faithfulness axiom is satisfied.

All axioms are satisfied.

**Uniqueness.** Let  $\nu'_{U, \mathcal{D}, \Pi}$  be any group-level data valuation method satisfying Axioms 1-5 from Definition 2. We have the following progressive arguments.

1. Consider the partition  $\Pi^* = \{\{k\}\}_{k \in [n]}$ , where each group is a singleton. When Axioms 1-4 (null player, symmetry, linearity, and efficiency) are applied to the groups  $\{k\} \in \Pi^*$ , they correspond

to the standard axioms for individual player valuations. The Shapley value  $SV(k)$  is the unique valuation satisfying these axioms for individual players. Thus, for any  $k \in [n]$ , we must have  $\nu'_{U, \mathcal{D}, \Pi^*}(\{k\}) = SV(k)$ .

2. For any non-empty subset  $S \subseteq [n]$ , consider the specific partition  $\Pi_S = \{S\} \cup \{\{j\}\}_{j \in [n] \setminus S}$ . For any singleton group  $\{j\}$  where  $j \in [n] \setminus S$ , note that  $\{j\} \in \Pi_S$  and also  $\{j\} \in \Pi^*$ . By Axiom 5 (faithfulness):

$$\nu'_{U, \mathcal{D}, \Pi_S}(\{j\}) = \nu'_{U, \mathcal{D}, \Pi^*}(\{j\}).$$

Using the result from step 1,  $\nu'_{U, \mathcal{D}, \Pi^*}(\{j\}) = SV(j)$ . So, for  $j \in [n] \setminus S$ :

$$\nu'_{U, \mathcal{D}, \Pi_S}(\{j\}) = SV(j).$$

Now, apply Axiom 4 (efficiency) to the partition  $\Pi_S$ :

$$\nu'_{U, \mathcal{D}, \Pi_S}(S) + \sum_{j \in [n] \setminus S} \nu'_{U, \mathcal{D}, \Pi_S}(\{j\}) = U([n]).$$

Substituting  $\nu'_{U, \mathcal{D}, \Pi_S}(\{j\}) = SV(j)$  for  $j \in [n] \setminus S$ :

$$\nu'_{U, \mathcal{D}, \Pi_S}(S) + \sum_{j \in [n] \setminus S} SV(j) = U([n]).$$

By the efficiency of individual Shapley values,  $U([n]) = \sum_{k \in [n]} SV(k)$ . Therefore:

$$\nu'_{U, \mathcal{D}, \Pi_S}(S) = \sum_{k \in [n]} SV(k) - \sum_{j \in [n] \setminus S} SV(j) = \sum_{i \in S} SV(i).$$

3. Now, consider any arbitrary partition  $\Pi$  of  $[n]$  such that  $S \in \Pi$ . Since  $S \in \Pi$  and  $S \in \Pi_S$ , by Axiom 5 (faithfulness):

$$\nu'_{U, \mathcal{D}, \Pi}(S) = \nu'_{U, \mathcal{D}, \Pi_S}(S).$$

Substituting the result from step 2:

$$\nu'_{U, \mathcal{D}, \Pi}(S) = \sum_{i \in S} SV(i).$$

Since this holds for any valuation  $\nu'$  satisfying the axioms and for any  $S \in \Pi$ , the valuation method is unique and given by the sum of individual Shapley values.  $\square$

### A.3 Proof of Lemma 1

*Proof of Lemma 1.* By definition,  $\text{FGSV}(S_0) = \sum_{i \in S_0} SV(i)$ . Using the standard definition of the individual Shapley value  $SV(i)$  from (1):

$$\text{FGSV}(S_0) = \sum_{i \in S_0} \sum_{S \subseteq [n] \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} [U(S \cup \{i\}) - U(S)]. \quad (5)$$

The definition (5) shows that  $\text{FGSV}(S_0)$  is a linear combination of the utility values  $U(S)$  for  $S \subseteq [n]$ . We now determine the coefficient  $C(S)$  for a specific subset  $S$  with cardinality  $s = |S|$ . A term  $U(S)$  appears positively in the sum for  $SV(i)$  if  $i \in S_0 \cap S$ , and negatively if  $i \in S_0 \setminus S$ . The respective Shapley weights are  $\frac{(s-1)!(n-s)!}{n!}$  and  $\frac{s!(n-s-1)!}{n!}$ . Summing over  $i \in S_0$ , the coefficient  $C(S)$  is therefore:

$$C(S) = |S_0 \cap S| \frac{(s-1)!(n-s)!}{n!} - |S_0 \setminus S| \frac{s!(n-s-1)!}{n!}.$$

Let  $s_1 = |S_0 \cap S|$ . Then  $|S_0 \setminus S| = s_0 - s_1$ . Substituting  $s = |S|$  and simplifying the factorials (for  $0 < s < n$ ):

$$\begin{aligned} C(S) &= s_1 \frac{(s-1)!(n-s)!}{n!} - (s_0 - s_1) \frac{s!(n-s-1)!}{n!} \\ &= \frac{s!(n-s)!}{n!} \left[ \frac{s_1}{s} - \frac{s_0 - s_1}{n-s} \right] \\ &= \frac{1}{\binom{n}{s}} \left[ \frac{s_1(n-s) - (s_0 - s_1)s}{s(n-s)} \right] = \frac{s_1 n - s_0 s}{s(n-s)} \frac{1}{\binom{n}{s}}. \end{aligned}$$



This coefficient depends only on  $s = |S|$  and  $s_1 = |S_0 \cap S|$ . We rewrite  $\text{FGSV}(S_0) = \sum_{S \subseteq [n]} C(S)U(S)$  by grouping terms with the same  $s$  and  $s_1$ . We handle the edge cases  $s = 0$  ( $S = \emptyset$ ) and  $s = n$  ( $S = [n]$ ) separately. For  $S = \emptyset$  ( $s = 0$ ),  $s_1 = 0$ , the coefficient is  $C(\emptyset) = -s_0/n$ . For  $S = [n]$  ( $s = n$ ),  $s_1 = s_0$ , the coefficient is  $C([n]) = s_0/n$ . So we can write:

$$\begin{aligned} \text{FGSV}(S_0) &= C([n])U([n]) + C(\emptyset)U(\emptyset) + \sum_{s=1}^{n-1} \sum_{\substack{S \subseteq [n] \\ |S|=s}} C(S)U(S) \\ &= \frac{s_0}{n} [U([n]) - U(\emptyset)] + \sum_{s=1}^{n-1} \sum_{\substack{S \subseteq [n] \\ |S|=s}} \left( \frac{s_1 n - s_0 s}{s(n-s)} \frac{1}{\binom{n}{s}} \right) U(S), \end{aligned}$$

where  $s_1 = |S \cap S_0|$  within the inner sum. Now, we focus on handling

$$\mathcal{T} := \sum_{s=1}^{n-1} \sum_{\substack{S \subseteq [n] \\ |S|=s}} \left( \frac{s_1 n - s_0 s}{s(n-s)} \frac{1}{\binom{n}{s}} \right) U(S).$$

By grouping the inner sum by the value of  $s_1 = |S \cap S_0|$  for each fixed size  $s$ , we have:

$$\mathcal{T} = \sum_{s=1}^{n-1} \sum_{s_1=\max\{0, s+s_0-n\}}^{\min\{s, s_0\}} \left( \frac{s_1 n - s_0 s}{s(n-s)} \frac{1}{\binom{n}{s}} \right) \sum_{\substack{S: |S|=s \\ |S \cap S_0|=s_1}} U(S).$$

Using the definition  $\mu\left(\frac{s_1}{s}; s, s_0, n\right) = \frac{\sum_{S: |S|=s, |S \cap S_0|=s_1} U(S)}{\binom{s_0}{s_1} \binom{n-s_0}{s-s_1}}$  (from (6)), we have:

$$\mathcal{T} = \sum_{s=1}^{n-1} \sum_{s_1} \frac{s_1 n - s_0 s}{s(n-s)} \frac{\binom{s_0}{s_1} \binom{n-s_0}{s-s_1}}{\binom{n}{s}} \mu\left(\frac{s_1}{s}; s, s_0, n\right).$$

Recognizing the hypergeometric probability  $\mathbb{P}_{\mathbf{s}_1 \sim \mathcal{HG}(n, s_0, s)}(\mathbf{s}_1 = s_1) = \frac{\binom{s_0}{s_1} \binom{n-s_0}{s-s_1}}{\binom{n}{s}}$ :

$$\begin{aligned} \mathcal{T} &= \sum_{s=1}^{n-1} \sum_{s_1} \mathbb{P}(\mathbf{s}_1 = s_1) \frac{n(s_1 - ss_0/n)}{s(n-s)} \mu\left(\frac{s_1}{s}; s, s_0, n\right) \\ &= \sum_{s=1}^{n-1} \mathbb{E}_{\mathbf{s}_1 \sim \mathcal{HG}(n, s_0, s)} \left[ \frac{n}{s(n-s)} \left( \mathbf{s}_1 - \frac{ss_0}{n} \right) \mu\left(\frac{\mathbf{s}_1}{s}; s, s_0, n\right) \right] \\ &= \sum_{s=1}^{n-1} \mathcal{T}(s), \end{aligned}$$

where  $\mathcal{T}(s)$  matches the definition (8). Combining the parts, we get  $\text{FGSV}(S_0) = \frac{s_0}{n} [U([n]) - U(\emptyset)] + \sum_{s=1}^{n-1} \mathcal{T}(s)$ , proving the lemma.  $\square$

#### A.4 Proof of Theorem 2

Before presenting the proof of Theorem 2, we establish two lemmas that demonstrate how Assumption 2 implies a smoothness condition on the function  $\mu(\cdot)$ . The first lemma shows that this assumption yields a bound on the second-order finite difference of  $\mu$ .

**Lemma 5.** *Under Assumption 2, the following second-order difference bound holds for  $\mu\left(\frac{s_1}{s}; s, s_0, n\right)$ :*

$$\left| \mu\left(\frac{s_1+1}{s}; s, s_0, n\right) - 2\mu\left(\frac{s_1}{s}; s, s_0, n\right) + \mu\left(\frac{s_1-1}{s}; s, s_0, n\right) \right| \leq \frac{C}{s^{\frac{3}{2}+v}}. \quad (6)$$

The proof of Lemma 5 is in Appendix A.4.1.

Next, we show that this second-order difference bound implies a first-order approximation error bound – a discrete analogue of Taylor’s theorem. To formalize this, we define the following continuity-extended versions of a discrete function and its first-order derivative.

**Definition 1** (Continuity-extended function). *Let  $f_s(x)$  be a function defined on rational points of the form  $x = \frac{s_1}{s}$  for integers  $s_1 \in \{0, 1, \dots, s\}$ . The continuity-extended function  $\tilde{f}_s(x)$  is defined as the linear interpolation:*

$$\tilde{f}_s(x) = \begin{cases} (sx - \lfloor sx \rfloor) \cdot f_s\left(\frac{\lfloor sx \rfloor + 1}{s}\right) + (\lfloor sx \rfloor + 1 - sx) \cdot f_s\left(\frac{\lfloor sx \rfloor}{s}\right), & x \in [0, 1), \\ f_s(1), & x = 1. \end{cases}$$

Intuitively,  $\tilde{f}_s$  linearly interpolates  $f_s$  between adjacent grid points. Notice that  $\tilde{f}_s$  is indeed smooth in between, thus  $\tilde{f}_s'$  is well-defined there. When taking the derivative of  $\tilde{f}_s$  w.r.t.  $x$ , notice that  $\lfloor sx \rfloor$  is constant between grid points, thus differentiate to zero. This naturally leads to the following definition.

**Definition 2** (Continuity-extended first-order derivative). *Let  $f_s(x)$  be as above. The continuity-extended first-order derivative  $\tilde{f}_s'(x)$  is defined as:*

$$\tilde{f}_s'(x) = s \cdot \left[ f_s\left(\frac{\lfloor sx \rfloor + 1}{s}\right) - f_s\left(\frac{\lfloor sx \rfloor}{s}\right) \right], \quad x \in (0, 1).$$

**Lemma 6.** *Let  $\{f_s(x)\}_{s=1}^\infty$  be a sequence of functions defined on rational points  $x = \frac{s_1}{s}$  with  $s_1 \in \{0, 1, \dots, s\}$ . Suppose  $f_s(x)$  satisfies the discrete second-order difference bound:*

$$\left| f_s\left(\frac{s_1+1}{s}\right) - 2f_s\left(\frac{s_1}{s}\right) + f_s\left(\frac{s_1-1}{s}\right) \right| \leq \frac{C}{s^{\frac{3}{2}+v}}, \quad 1 \leq s_1 < s-1, \quad (7)$$

for constants  $C > 0$  and  $v > 0$ . Then for all  $s_1, x$  with  $0 < \frac{s_1}{s}, x < 1$ , the following first-order approximation holds:

$$\left| f_s\left(\frac{s_1}{s}\right) - \tilde{f}_s(x) - \tilde{f}_s'(x) \cdot \left(\frac{s_1}{s} - x\right) \right| \leq C s^{\frac{1}{2}-v} \left(\frac{s_1}{s} - x\right)^2. \quad (8)$$

To intuitively understand Lemma 6, notice that by definition, we have

$$f_s\left(\frac{s_1}{s}\right) - \tilde{f}_s(x) - \tilde{f}_s'(x) \cdot \left(\frac{s_1}{s} - x\right) = 0, \quad \text{for } x \in [s_1/s, (s_1+1)/s]. \quad (9)$$

Then in view of (7), it is not difficult to understand that the error bound on the RHS of (8) is the consequence of bounding telescoping sums that eventually reduces the problem to the case of (9), summing up the approximation errors along the way.

The formal proof of Lemma 6 is in Appendix A.4.1.

We now proceed to prove Theorem 2.

*Proof of Theorem 2.* The core idea is to approximate the function  $\mu(\frac{s_1}{s}; s, s_0, n)$  inside the expectation defining  $\mathcal{T}(s)$  using a first-order Taylor-like expansion around the mean proportion  $\alpha_0 = s_0/n$ . The validity of this expansion relies on the smoothness properties derived from Assumption 2.

Let  $\mu_s(x) := \mu(x; s, s_0, n)$  denote the function  $\mu$  for fixed  $s, s_0, n$ , where  $x = s_1/s$  is the proportion of intersection. By Lemma 5, Assumption 2 ensures that  $\mu_s$  satisfies the second-order difference bound given in (6). We can then apply Lemma 6 (First-Order Approximation Bound) with  $f_s \rightarrow \mu_s$ , the evaluation point  $x_0 = s_1/s$ , and the expansion point  $x = \alpha_0$ , and get the following approximation for  $\mu$ :

$$\mu\left(\frac{s_1}{s}; s, s_0, n\right) = \tilde{\mu}(\alpha_0; s, s_0, n) + \tilde{\mu}'(\alpha_0; s, s_0, n) \cdot \left(\frac{s_1}{s} - \alpha_0\right) + R\left(\frac{s_1}{s}; s, s_0, n\right),$$

where  $\tilde{\mu}$  and  $\tilde{\mu}'$  are the continuity-extended function and its first-order difference defined in Definitions 1 and 2, respectively. The remainder term  $R$  satisfies the bound from Lemma 6 (Eq. (8)):

$$\left| R\left(\frac{s_1}{s}; s, s_0, n\right) \right| \leq C' s^{\frac{1}{2}-v} \left(\frac{s_1}{s} - \alpha_0\right)^2 = C' s^{-\frac{3}{2}-v} (s_1 - s\alpha_0)^2,$$

with  $C'$  depending on the constant  $C$  from Assumption 2.

Now, we substitute this expansion of  $\mu$  into the definition of  $\mathcal{T}(s)$  (Eq. (8)):

$$\begin{aligned} \mathcal{T}(s) &= \mathbb{E}_{\mathbf{s}_1 \sim \mathcal{HG}(n, s_0, s)} \left[ \frac{n}{s(n-s)} (\mathbf{s}_1 - s\alpha_0) \mu\left(\frac{\mathbf{s}_1}{s}; s, s_0, n\right) \right] \\ &= \mathbb{E}_{\mathbf{s}_1} \left[ \frac{n}{s(n-s)} (\mathbf{s}_1 - s\alpha_0) \left( \tilde{\mu}(\alpha_0; s, s_0, n) + \tilde{\mu}'(\alpha_0; s, s_0, n) \left(\frac{\mathbf{s}_1}{s} - \alpha_0\right) + R\left(\frac{\mathbf{s}_1}{s}; s, s_0, n\right) \right) \right]. \end{aligned}$$

Using the linearity of expectation, we decompose  $\mathcal{T}(s)$  into three terms:

$$\begin{aligned} \mathcal{T}(s) &= \underbrace{\tilde{\mu}(\alpha_0; s, s_0, n) \mathbb{E}_{\mathbf{s}_1} \left[ \frac{n}{s(n-s)} (\mathbf{s}_1 - s\alpha_0) \right]}_{\text{Term I}} \\ &\quad + \underbrace{\tilde{\mu}'(\alpha_0; s, s_0, n) \cdot \mathbb{E}_{\mathbf{s}_1} \left[ \frac{n}{s(n-s)} (\mathbf{s}_1 - s\alpha_0) \left(\frac{\mathbf{s}_1}{s} - \alpha_0\right) \right]}_{\text{Term II}} \\ &\quad + \underbrace{\mathbb{E}_{\mathbf{s}_1} \left[ \frac{n}{s(n-s)} (\mathbf{s}_1 - s\alpha_0) R\left(\frac{\mathbf{s}_1}{s}; s, s_0, n\right) \right]}_{\text{Term III}}. \end{aligned}$$

**Analysis of Term I:** The random variable  $\mathbf{s}_1$  follows the Hypergeometric distribution  $\mathcal{HG}(n, s_0, s)$ , which has mean  $\mathbb{E}[\mathbf{s}_1] = ss_0/n = s\alpha_0$ . Therefore,  $\mathbb{E}[\mathbf{s}_1 - s\alpha_0] = 0$ , which implies Term I = 0.

**Analysis of Term II:** The expectation in Term II involves the second central moment (variance) of  $\mathbf{s}_1$ :

$$\mathbb{E}_{\mathbf{s}_1} \left[ (\mathbf{s}_1 - s\alpha_0) \left(\frac{\mathbf{s}_1}{s} - \alpha_0\right) \right] = \mathbb{E}_{\mathbf{s}_1} \left[ \frac{1}{s} (\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1)^2 \right] = \frac{1}{s} \text{Var}(\mathbf{s}_1).$$

The variance of  $\mathcal{HG}(n, s_0, s)$  is  $\text{Var}(\mathbf{s}_1) = s\alpha_0(1 - \alpha_0) \frac{n-s}{n-1}$ . Substituting this into Term II:

$$\begin{aligned} \text{Term II} &= \tilde{\mu}'(\alpha_0; s, s_0, n) \cdot \frac{n}{s(n-s)} \cdot \frac{1}{s} \text{Var}(\mathbf{s}_1) \\ &= \tilde{\mu}'(\alpha_0; s, s_0, n) \cdot \frac{n}{s^2(n-s)} \cdot \left( s\alpha_0(1 - \alpha_0) \frac{n-s}{n-1} \right) \\ &= \tilde{\mu}'(\alpha_0; s, s_0, n) \cdot \frac{n\alpha_0(1 - \alpha_0)}{s(n-1)}. \end{aligned}$$

Now, substitute the definition of  $\tilde{\mu}'$  from Definition 2 evaluated at  $x = \alpha_0 = s_0/n$ . Let  $s_1^* = \lfloor s\alpha_0 \rfloor$ . Then:

$$\tilde{\mu}'(\alpha_0; s, s_0, n) = s \left[ \mu\left(\frac{s_1^* + 1}{s}; s, s_0, n\right) - \mu\left(\frac{s_1^*}{s}; s, s_0, n\right) \right] = s \Delta\mu\left(\frac{s_1^*}{s}; s, s_0, n\right).$$

Substituting this into the expression for Term II yields:

$$\text{Term II} = \left( s \Delta\mu\left(\frac{s_1^*}{s}; s, s_0, n\right) \right) \cdot \frac{n\alpha_0(1 - \alpha_0)}{s(n-1)} = \frac{n}{n-1} \alpha_0(1 - \alpha_0) \Delta\mu\left(\frac{s_1^*}{s}; s, s_0, n\right).$$

**Analysis of Term III (Remainder Term):** We bound the absolute value using the bound on  $|R|$  from Lemma 6:

$$\begin{aligned}
|\text{Term III}| &= \left| \mathbb{E}_{\mathbf{s}_1} \left[ \frac{n}{s(n-s)} (\mathbf{s}_1 - s\alpha_0) R \left( \frac{\mathbf{s}_1}{s}; s, s_0, n \right) \right] \right| \\
&\leq \frac{n}{s(n-s)} \mathbb{E}_{\mathbf{s}_1} \left[ |\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1| \cdot \left| R \left( \frac{\mathbf{s}_1}{s}; s, s_0, n \right) \right| \right] \\
&\leq \frac{n}{s(n-s)} \mathbb{E}_{\mathbf{s}_1} \left[ |\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1| \cdot C s^{-\frac{3}{2}-v} (\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1)^2 \right] \\
&= \frac{nC}{s^{5/2+v}(n-s)} \mathbb{E}_{\mathbf{s}_1} [|\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1|^3].
\end{aligned}$$

Using Cauchy-Schwarz ( $\mathbb{E}[|X|^3] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[X^4]}$ ) where  $X = \mathbf{s}_1 - \mathbb{E}\mathbf{s}_1$ , and the fact that  $\text{Var}(\mathbf{s}_1) = O\left(\alpha_0(1-\alpha_0)\frac{s(n-s)}{n}\right)$  and Kurtosis( $\mathbf{s}_1$ ) =  $O\left(\frac{n}{s(n-s)\alpha_0(1-\alpha_0)}\right) = O\left(\frac{1}{\alpha_0(1-\alpha_0)}\right)$  for  $1 \leq s \leq n-1$ , we yields:

$$\begin{aligned}
\mathbb{E}_{\mathbf{s}_1} [|\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1|^3] &\lesssim \sqrt{\mathbb{E}_{\mathbf{s}_1} [|\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1|^2] \cdot \mathbb{E}_{\mathbf{s}_1} [|\mathbf{s}_1 - \mathbb{E}\mathbf{s}_1|^4]} \\
&\lesssim \sqrt{\text{Var}(\mathbf{s}_1) \cdot (\text{Kurtosis}(\mathbf{s}_1) + 3)(\text{Var}(\mathbf{s}_1))^2} \\
&\lesssim \alpha_0(1-\alpha_0) \left( \frac{s(n-s)}{n-1} \right)^{3/2}.
\end{aligned}$$

Substituting this into the bound for  $|\text{Term III}|$ :

$$\begin{aligned}
|\text{Term III}| &\lesssim \frac{n}{s^{5/2+v}(n-s)} \alpha_0(1-\alpha_0) \left( \frac{s(n-s)}{n-1} \right)^{3/2} \\
&\lesssim n\alpha_0(1-\alpha_0) \frac{(n-s)^{1/2}}{(n-1)^{3/2}} s^{-1-v} \\
&= O\left( \frac{n}{n-1} \alpha_0(1-\alpha_0) s^{-(1+v)} \right).
\end{aligned}$$

So, Term III contributes the  $O(s^{-(1+v)})$  error term, scaled by factors related to  $\alpha_0$  and  $n/(n-1)$ .

**Conclusion:** Combining Terms I=0, Term II, and the bound on Term III gives:

$$\begin{aligned}
\mathcal{T}(s) &= \frac{n}{n-1} \alpha_0(1-\alpha_0) \Delta\mu \left( \frac{s_1^*}{s}; s, s_0, n \right) + O\left( \frac{n}{n-1} \alpha_0(1-\alpha_0) s^{-(1+v)} \right) \\
&= \frac{n}{n-1} \alpha_0(1-\alpha_0) \left[ \Delta\mu \left( \frac{s_1^*}{s}; s, s_0, n \right) + O(s^{-(1+v)}) \right],
\end{aligned}$$

where the constant factor in the  $O(\cdot)$  term depends on  $C, C', v$ , but not on  $s$  and  $\alpha_0$ . This proves Theorem 2.  $\square$

#### A.4.1 Technical lemmas used in the proof of Theorem 2

**Lemma 5.** Under Assumption 2, the following second-order difference bound holds for  $\mu\left(\frac{s_1}{s}; s, s_0, n\right)$ :

$$\left| \mu\left(\frac{s_1+1}{s}; s, s_0, n\right) - 2\mu\left(\frac{s_1}{s}; s, s_0, n\right) + \mu\left(\frac{s_1-1}{s}; s, s_0, n\right) \right| \leq \frac{C}{s^{\frac{3}{2}+v}}. \quad (6)$$

*Proof of Lemma 5.* The core idea is to relate the second-order difference of  $\mu$  (which is an average of  $U$  values) to the average of the second-order difference quantity from Assumption 2. Specifically, we consider the stability term  $\Delta(S', z_1, z'_1, z_2, z'_2) := U(S' \cup \{z_1, z'_1\}) - U(S' \cup \{z'_1, z'_2\}) - U(S' \cup \{z_1, z_2\}) + U(S' \cup \{z_2, z'_2\})$ , which is bounded by  $C|S'|^{-(3/2+v)}$  by Assumption 2. We will construct a sum,  $\mathcal{M}(s_1, s_2)$ , by averaging  $\Delta(S', z_1, z'_1, z_2, z'_2)$  over appropriate base sets  $S'$  and points  $z_1, z'_1 \in S_0, z_2, z'_2 \notin S_0$ . We then show that this average value,  $\mathcal{M}(s_1, s_2)$  properly

normalized, exactly equals the second-order finite difference of  $\mu$  stated in the lemma. The bound on  $\mu$ 's second difference then follows from the bound on  $\Delta$ .

Let  $s_2 = s - s_1$ . We define the sum  $\mathcal{M}(s_1, s_2)$  over all possible configurations:

$$\mathcal{M}(s_1, s_2) := \sum_{\substack{S'_1 \subseteq S_0 \\ |S'_1|=s_1-1}} \sum_{\substack{S'_2 \subseteq S_0^c \\ |S'_2|=s_2-1}} \sum_{\substack{z_1, z'_1 \in S_0 \setminus S'_1 \\ z_1 \neq z'_1}} \sum_{\substack{z_2, z'_2 \in S_0^c \setminus S'_2 \\ z_2 \neq z'_2}} \Delta(S'_1 \cup S'_2, z_1, z'_1, z_2, z'_2), \quad (10)$$

where the sum is defined for  $s_1 \geq 1, s_2 \geq 1$  and requires  $|S_0 \setminus S'_1| \geq 2$  (i.e.,  $s_0 - (s_1 - 1) \geq 2 \implies s_1 \leq s_0 - 1$ ) and  $|S_0^c \setminus S'_2| \geq 2$  (i.e.,  $(n - s_0) - (s_2 - 1) \geq 2 \implies s_2 \leq n - s_0 - 1$ ).

Our goal now is to demonstrate that this aggregated sum  $\mathcal{M}(s_1, s_2)$ , when properly normalized, equals the second-order finite difference of the average utility function  $\mu$ . Specifically, we aim to prove the following identity:

$$\frac{\mathcal{M}(s_1, s_2)}{N_{\text{terms}}} = \mu\left(\frac{s_1 + 1}{s}; s, s_0, n\right) - 2\mu\left(\frac{s_1}{s}; s, s_0, n\right) + \mu\left(\frac{s_1 - 1}{s}; s, s_0, n\right), \quad (11)$$

where  $N_{\text{terms}}$  is the total number of  $\Delta(\cdot)$  terms summed in the definition of  $\mathcal{M}(s_1, s_2)$  (Eq. (10)). We establish this identity through the following combinatorial analysis.

Firstly, when the  $\Delta(\cdot)$  terms are expanded,  $\mathcal{M}(s_1, s_2)$  becomes a sum of individual  $U(S)$  terms. Each such set  $S$  has size  $s = s_1 + s_2$ . Furthermore, based on the structure of  $\Delta(S', z_1, z'_1, z_2, z'_2)$ , the number of elements from  $S_0$  in any  $S$  appearing with a non-zero coefficient must be  $s_1 + 1, s_1$ , or  $s_1 - 1$ . Let  $\mathcal{A}_{s'_1, s}$  denote the collection of all sets  $S \subseteq [n]$  such that  $|S| = s$  and  $|S \cap S_0| = s'_1$ . Due to the symmetric construction of  $\mathcal{M}(s_1, s_2)$ , all  $U(S)$  terms for  $S$  within the same collection  $\mathcal{A}_{s'_1, s}$  appear with the same net coefficient in the expansion of  $\mathcal{M}(s_1, s_2)$ .

Secondly, we determine these net coefficients. Let  $C(s'_1, s)$  be the coefficient for any  $U(S)$  where  $S \in \mathcal{A}_{s'_1, s}$ . A detailed combinatorial count, considering the base sets  $S' = S'_1 \cup S'_2$  and the ordered pairs  $(z_1, z'_1)$  and  $(z_2, z'_2)$  involved in the definition of  $\mathcal{M}(s_1, s_2)$ , yields the following coefficients:

- For sets  $S \in \mathcal{A}_{s_1+1, s}$  (meaning  $|S \cap S_0| = s_1 + 1$  and thus  $|S \cap S_0^c| = s - (s_1 + 1) = s_2 - 1$ , where  $s_2 = s - s_1$ ): These utility terms  $U(S)$  arise solely from the  $+U(S' \cup \{z_1, z'_1\})$  component in the expansion of  $\Delta(S', z_1, z'_1, z_2, z'_2)$  within the sum  $\mathcal{M}(s_1, s_2)$  defined in (10). For a fixed  $S = S_1 \cup S_2$  in this collection (with  $|S_1| = s_1 + 1$  and  $|S_2| = s_2 - 1$ ), the number of ways to form  $S$  via  $(S', z_1, z'_1, z_2, z'_2)$  requires choosing ordered pairs  $(z_1, z'_1) \in S_1, (z_2, z'_2) \in S_0^c \setminus S_2$ . The resulting coefficient is:

$$C(s_1 + 1, s) = (s_1 + 1)s_1 \times (n - s_0 - s_2 + 1)(n - s_0 - s_2).$$

- For  $S \in \mathcal{A}_{s_1, s}$  (which corresponds to  $s'_2 = s_2$ ): Terms arise from  $-U(S' \cup \{z_1, z_2\})$  and  $-U(S' \cup \{z'_1, z'_2\})$  in the expansion of  $\Delta(S', z_1, z'_1, z_2, z'_2)$ . For a fixed  $S = S_1 \cup S_2$  (with  $|S_1| = s_1, |S_2| = s_2$ ), the number of ways to form it via  $(S', z_1, z'_1, z_2, z'_2)$  requires choosing  $S_1 = S'_1 \setminus \{z_1\}, S_2 = S'_2 \setminus \{z_2\}, z'_1 \in S_0 \setminus S'_1, z'_2 \in S_0^c \setminus S'_2$ . Then the resulting coefficient is:

$$C(s_1, s) = s_1(n - s_0 - s_2) \times (s_0 - s_1)s_2.$$

- For  $S \in \mathcal{A}_{s_1-1, s}$  (which corresponds to  $s'_2 = s_2 + 1$ ): The term arises only from  $+U(S' \cup \{z_2, z'_2\})$ . With similar argument for  $\mathcal{A}_{s_1+1, s}$ , the resulting coefficient is:

$$C(s_1 - 1, s) = (s_0 - s_1 + 1)(s_0 - s_1) \times (s - s_1 + 1)(s - s_1).$$

Thirdly, we determine the total number of  $\Delta(\cdot)$  terms in the sum  $\mathcal{M}(s_1, s_2)$ . The number of ways to choose the base set  $S' = S'_1 \cup S'_2$  (with  $|S'_1| = s_1 - 1, |S'_2| = s_2 - 1$ ) is  $N_{S'} = \binom{s_0}{s_1-1} \binom{n-s_0}{s_2-1}$ . For each  $S'$ , the number of ordered pairs  $(z_1, z'_1)$  with  $z_1 \neq z'_1$  from  $S_0 \setminus S'_1$  is  $N_{z1} = P(s_0 - s_1 + 1, 2) = (s_0 - s_1 + 1)(s_0 - s_1)$ . The number of ordered pairs  $(z_2, z'_2)$  with  $z_2 \neq z'_2$  from  $S_0^c \setminus S'_2$  is  $N_{z2} = P(n - s_0 - s_2 + 1, 2) = (n - s_0 - s_2 + 1)(n - s_0 - s_2)$ . Therefore, the total number of terms is:

$$N_{\text{terms}} = N_{S'} \times N_{z1} \times N_{z2} = \binom{s_0}{s_1-1} \binom{n-s_0}{s_2-1} (s_0 - s_1 + 1)(s_0 - s_1)(n - s_0 - s_2 + 1)(n - s_0 - s_2).$$

Gathering these observations, the crucial step relies on the fact that the derived coefficients  $C(\cdot)$  and the normalization  $N_{\text{terms}}$ , when combined with the number of sets in each collection  $|\mathcal{A}_{s'_1, s}| = \binom{s_0}{s'_1} \binom{n-s_0}{s-s'_1}$ , simplify exactly as follows:

$$\begin{aligned} \frac{\mathcal{M}(s_1, s_2)}{N_{\text{terms}}} &= \frac{C(s_1 + 1, s)}{N_{\text{terms}}} \sum_{S \in \mathcal{A}_{s_1+1, s}} U(S) + \frac{C(s_1, s)}{N_{\text{terms}}} \sum_{S \in \mathcal{A}_{s_1, s}} U(S) + \frac{C(s_1 - 1, s)}{N_{\text{terms}}} \sum_{S \in \mathcal{A}_{s_1-1, s}} U(S) \\ &= +1 \cdot \frac{\sum_{S \in \mathcal{A}_{s_1+1, s}} U(S)}{|\mathcal{A}_{s_1+1, s}|} - 2 \cdot \frac{\sum_{S \in \mathcal{A}_{s_1, s}} U(S)}{|\mathcal{A}_{s_1, s}|} + 1 \cdot \frac{\sum_{S \in \mathcal{A}_{s_1-1, s}} U(S)}{|\mathcal{A}_{s_1-1, s}|} \\ &= \mu \left( \frac{s_1 + 1}{s}; s, s_0, n \right) - 2\mu \left( \frac{s_1}{s}; s, s_0, n \right) + \mu \left( \frac{s_1 - 1}{s}; s, s_0, n \right). \end{aligned} \quad (12)$$

This establishes the identity (11) relating the average  $\Delta$  value to the second difference of  $\mu$ .

Now, we bound the absolute value of the left-hand side of (6). We have

$$\begin{aligned} |\text{RHS of (12)}| &= \left| \frac{1}{N_{\text{terms}}} \sum_{S', z_1, z'_1, z_2, z'_2} \Delta(S', z_1, z'_1, z_2, z'_2) \right| \\ &\leq \frac{1}{N_{\text{terms}}} \sum_{S', z_1, z'_1, z_2, z'_2} |\Delta(S', z_1, z'_1, z_2, z'_2)| \quad (\text{By Triangle Inequality}) \\ &\leq \max_{S', z_1, z'_1, z_2, z'_2} |\Delta(S', z_1, z'_1, z_2, z'_2)|. \end{aligned}$$

By Assumption 2, for any base set  $S'$  (of size  $s - 2$ ) and points  $z_1, z'_1, z_2, z'_2$ :

$$|\Delta(S', z_1, z'_1, z_2, z'_2)| \leq C|S'|^{-(3/2+v)} = C(s-2)^{-(3/2+v)}.$$

Therefore,

$$|\text{RHS of (12)}| \leq C(s-2)^{-(3/2+v)}.$$

For  $s \geq 2$ ,  $(s-2)^{-(3/2+v)}$  is of the order  $O(s^{-(3/2+v)})$ . We can absorb the constant factor difference between  $(s-2)^{-k}$  and  $s^{-k}$  into a modified constant  $C$  (or  $C'$ ), yielding the desired bound:

$$\left| \mu \left( \frac{s_1 + 1}{s}; s, s_0, n \right) - 2\mu \left( \frac{s_1}{s}; s, s_0, n \right) + \mu \left( \frac{s_1 - 1}{s}; s, s_0, n \right) \right| \leq \frac{C'}{s^{3/2+v}}.$$

This completes the proof.  $\square$

**Lemma 6.** Let  $\{f_s(x)\}_{s=1}^\infty$  be a sequence of functions defined on rational points  $x = \frac{s_1}{s}$  with  $s_1 \in \{0, 1, \dots, s\}$ . Suppose  $f_s(x)$  satisfies the discrete second-order difference bound:

$$\left| f_s \left( \frac{s_1 + 1}{s} \right) - 2f_s \left( \frac{s_1}{s} \right) + f_s \left( \frac{s_1 - 1}{s} \right) \right| \leq \frac{C}{s^{\frac{3}{2}+v}}, \quad 1 \leq s_1 < s-1, \quad (7)$$

for constants  $C > 0$  and  $v > 0$ . Then for all  $s_1, x$  with  $0 < \frac{s_1}{s}, x < 1$ , the following first-order approximation holds:

$$\left| f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x) - \tilde{f}'_s(x) \cdot \left( \frac{s_1}{s} - x \right) \right| \leq C s^{\frac{1}{2}-v} \left( \frac{s_1}{s} - x \right)^2. \quad (8)$$

*Proof of Lemma 6.* Let  $\tilde{f}'_s(y)$  be the continuity-extended first-order difference as defined in Definition 2. For any integer  $t$  such that  $0 \leq t < s$ ,  $\tilde{f}'_s(\frac{t}{s}) = s[f_s(\frac{t+1}{s}) - f_s(\frac{t}{s})]$ . The discrete second-order difference bound (7) states that for  $1 \leq t \leq s-1$ :

$$\left| f_s \left( \frac{t+1}{s} \right) - 2f_s \left( \frac{t}{s} \right) + f_s \left( \frac{t-1}{s} \right) \right| \leq \frac{C}{s^{\frac{3}{2}+v}}.$$

Consider the difference of  $\tilde{f}'_s$  at consecutive grid points  $t/s$  and  $(t-1)/s$  for  $1 \leq t < s$ :

$$\begin{aligned} \left| \tilde{f}'_s \left( \frac{t}{s} \right) - \tilde{f}'_s \left( \frac{t-1}{s} \right) \right| &= \left| s \left[ f_s \left( \frac{t+1}{s} \right) - f_s \left( \frac{t}{s} \right) \right] - s \left[ f_s \left( \frac{t}{s} \right) - f_s \left( \frac{t-1}{s} \right) \right] \right| \\ &= s \left| f_s \left( \frac{t+1}{s} \right) - 2f_s \left( \frac{t}{s} \right) + f_s \left( \frac{t-1}{s} \right) \right| \\ &\leq s \cdot \frac{C}{s^{\frac{3}{2}+v}} = \frac{C}{s^{\frac{1}{2}+v}}. \quad (\text{This holds for } 1 \leq t \leq s-1). \end{aligned}$$

This implies a Lipschitz-like condition for  $\tilde{f}'_s$  between grid points. For any two grid points  $x_a = a/s$  and  $x_b = b/s$  (assume  $a < b$  without loss of generality):

$$\begin{aligned} \left| \tilde{f}'_s(x_b) - \tilde{f}'_s(x_a) \right| &= \left| \sum_{j=a}^{b-1} \left( \tilde{f}'_s \left( \frac{j+1}{s} \right) - \tilde{f}'_s \left( \frac{j}{s} \right) \right) \right| \\ &\leq \sum_{j=a}^{b-1} \left| \tilde{f}'_s \left( \frac{j+1}{s} \right) - \tilde{f}'_s \left( \frac{j}{s} \right) \right| \leq \sum_{j=a}^{b-1} \frac{C}{s^{\frac{1}{2}+v}} = (b-a) \frac{C}{s^{\frac{1}{2}+v}}. \end{aligned}$$

So, for any two grid points  $x_i = i/s, x_j = j/s$ ,  $|\tilde{f}'_s(x_i) - \tilde{f}'_s(x_j)| \leq \frac{C}{s^{\frac{1}{2}+v}} |i - j|$ . Since  $\tilde{f}'_s(x)$  is piecewise constant between grid points (equal to  $\tilde{f}'_s(\lfloor sx \rfloor / s)$ ), for a general  $x \in [0, 1)$  and a grid point  $t/s$ :

$$\left| \tilde{f}'_s \left( \frac{t}{s} \right) - \tilde{f}'_s(x) \right| = \left| \tilde{f}'_s \left( \frac{t}{s} \right) - \tilde{f}'_s \left( \frac{\lfloor sx \rfloor}{s} \right) \right| \leq \frac{C}{s^{\frac{1}{2}+v}} |t - \lfloor sx \rfloor|. \quad (13)$$

We want to bound  $\left| f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x) - \tilde{f}'_s(x) \cdot \left( \frac{s_1}{s} - x \right) \right|$ . Consider the case  $\frac{s_1}{s} > x$ . The term  $f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x)$  can be written as a sum of first differences plus a boundary term:

$$\begin{aligned} f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x) &= \left( f_s \left( \frac{s_1}{s} \right) - f_s \left( \frac{s_1-1}{s} \right) + \dots + f_s \left( \frac{\lfloor sx \rfloor + 1}{s} \right) - f_s \left( \frac{\lfloor sx \rfloor}{s} \right) \right) - \left( \tilde{f}_s(x) - f_s \left( \frac{\lfloor sx \rfloor}{s} \right) \right) \\ &= \sum_{t=\lfloor sx \rfloor}^{s_1-1} \left( f_s \left( \frac{t+1}{s} \right) - f_s \left( \frac{t}{s} \right) \right) - \left( (sx - \lfloor sx \rfloor) \left( f_s \left( \frac{\lfloor sx \rfloor + 1}{s} \right) - f_s \left( \frac{\lfloor sx \rfloor}{s} \right) \right) \right) \quad (\text{by definition of } \tilde{f}_s) \\ &= \frac{1}{s} \sum_{t=\lfloor sx \rfloor}^{s_1-1} \tilde{f}'_s \left( \frac{t}{s} \right) - (sx - \lfloor sx \rfloor) \frac{1}{s} \tilde{f}'_s \left( \frac{\lfloor sx \rfloor}{s} \right) \\ &= \frac{1}{s} \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \tilde{f}'_s \left( \frac{t}{s} \right) + (\lfloor sx \rfloor + 1 - sx) \frac{1}{s} \tilde{f}'_s \left( \frac{\lfloor sx \rfloor}{s} \right) \\ &= \frac{1}{s} \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \tilde{f}'_s \left( \frac{t}{s} \right) + (\lfloor sx \rfloor + 1 - sx) \frac{1}{s} \tilde{f}'_s(x) \quad (\text{by definition of } \tilde{f}'_s) \end{aligned}$$

The expression to bound is:

$$\begin{aligned} f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x) - \tilde{f}'_s(x) \cdot \left( \frac{s_1}{s} - x \right) &= \left( \sum_{t=\lfloor sx \rfloor}^{s_1-1} \left[ f_s \left( \frac{t+1}{s} \right) - f_s \left( \frac{t}{s} \right) \right] + f_s \left( \frac{\lfloor sx \rfloor}{s} \right) \right) - \tilde{f}_s(x) - \tilde{f}'_s(x) \left( \frac{s_1}{s} - x \right). \end{aligned}$$

Therefore:

$$\begin{aligned}
& \left| f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x) - \tilde{f}'_s(x) \cdot \left( \frac{s_1}{s} - x \right) \right| \\
&= \left| \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \frac{1}{s} \tilde{f}'_s \left( \frac{t}{s} \right) + (\lfloor sx \rfloor + 1 - sx) \frac{1}{s} \tilde{f}'_s(x) - \left( \frac{s_1}{s} - x \right) \cdot \tilde{f}'_s(x) \right| \\
&= \left| \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \frac{1}{s} \tilde{f}'_s \left( \frac{t}{s} \right) - \left( \frac{s_1}{s} - \frac{\lfloor sx \rfloor + 1}{s} \right) \cdot \tilde{f}'_s(x) \right| \\
&= \left| \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \frac{1}{s} \tilde{f}'_s \left( \frac{t}{s} \right) - \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \frac{1}{s} \tilde{f}'_s(x) \right| \\
&= \frac{1}{s} \left| \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \left( \tilde{f}'_s \left( \frac{t}{s} \right) - \tilde{f}'_s(x) \right) \right| \\
&\leq \frac{1}{s} \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \left| \tilde{f}'_s \left( \frac{t}{s} \right) - \tilde{f}'_s(x) \right|.
\end{aligned}$$

Using the Lipschitz-like property from (13):

$$\begin{aligned}
\left| f_s \left( \frac{s_1}{s} \right) - \tilde{f}_s(x) - \tilde{f}'_s(x) \cdot \left( \frac{s_1}{s} - x \right) \right| &\leq \frac{1}{s} \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} \frac{C}{s^{\frac{1}{2}+v}} |t - \lfloor sx \rfloor| \\
&\leq \frac{C}{s^{\frac{3}{2}+v}} \sum_{t=\lfloor sx \rfloor + 1}^{s_1-1} |t - \lfloor sx \rfloor| \\
&\lesssim \frac{1}{s^{\frac{3}{2}+v}} (s_1 - sx)^2 \\
&\lesssim s^{\frac{1}{2}-v} \left( \frac{s_1}{s} - x \right)^2.
\end{aligned}$$

The case  $s_1 < sx$  follows similarly.  $\square$

### A.5 Proof of Theorem 3

Algorithm 1 employs a threshold,  $\bar{s}$ , to choose between two methods for estimating  $\mathcal{T}(s)$ :

- For  $s < \bar{s}$ , compute  $\hat{\mathcal{T}}(s)$  by directly approximating its definition formula (8), where each  $\mu$  term is estimated by subsampling according to its average form, as in (10).
- For  $s \geq \bar{s}$ , compute  $\hat{\mathcal{T}}(s)$  using the fast approximation formula (9), as per Theorem 2; the  $\Delta\mu$  term in this formula can be efficiently approximated by the paired Monte Carlo estimator elaborated in (11).

To analyze the overall error and derive the computational complexity, we first establish Lemma 7. This lemma bounds the total approximation error  $|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}|$  by combining concentration bounds for the Monte Carlo estimators  $\hat{\mu}_{m_1}$  and  $\hat{\Delta\mu}_{m_2}$  with the approximation error from Theorem 2.

**Lemma 7.** (Error Bound for  $\hat{\mathcal{T}}_{\text{sum}}$ ) Let  $\mathcal{T}_{\text{sum}} = \sum_{s=1}^{n-1} \mathcal{T}(s)$  and  $\hat{\mathcal{T}}_{\text{sum}} = \sum_{s=1}^{n-1} \hat{\mathcal{T}}(s)$  be the estimate computed by Algorithm 1 using threshold  $\bar{s}$  and sample sizes  $m_1, m_2$ . Under Assumptions 1 and 2, and assuming that the utility function  $U$  is  $\beta(s)$ -deletion stable, for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$ :

$$|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \lesssim \bar{s} \sqrt{\frac{\log(n/\delta)}{m_1}} + \alpha_0(1 - \alpha_0) \sqrt{\frac{\log(n/\delta)}{m_2}} \sum_{s=\bar{s}}^{n-1} \beta(s) + \alpha_0(1 - \alpha_0) O(\bar{s}^{-v}).$$



Building on this error bound, Lemma 8 establishes the computational cost for achieving an  $(\epsilon, \delta)$ -approximation of the sum  $\mathcal{T}_{\text{sum}}$ .

**Lemma 8.** (*Computational Cost for  $(\epsilon, \delta)$ -Approximation of  $\mathcal{T}_{\text{sum}}$* ) To achieve an  $(\epsilon, \delta)$ -approximation of  $\mathcal{T}_{\text{sum}} = \sum_{s=1}^{n-1} \mathcal{T}(s)$  (i.e., ensuring  $|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \leq \epsilon$  with probability at least  $1 - \delta$ ) using Algorithm 1, with parameters  $m_1, m_2, \bar{s}$  chosen optimally based on Lemma 7, the total number of utility function evaluations is:

$$O \left( \epsilon^{-\frac{4+2v}{v}} \log(n/\delta) + n \left[ 1 + \epsilon^{-2} (\alpha_0(1 - \alpha_0))^2 \left( \sum_{s=\bar{s}}^{n-1} \beta(s) \right)^2 \log(n/\delta) \right] \right).$$

With the computational cost for approximating  $\mathcal{T}_{\text{sum}} = \sum_{s=1}^{n-1} \mathcal{T}(s)$  established in Lemma 8, we are now equipped to prove our main complexity result for estimating  $\text{FGSV}(S_0)$ . Theorem 3 specifies this complexity under the condition that the deletion stability parameter  $\beta(s)$  decays as  $O(1/s)$ .

*Proof of Theorem 3.* By Lemma 1 in the main paper, we have  $\text{FGSV}(S_0) = G_0 + \mathcal{T}_{\text{sum}}$ , where  $G_0 = \frac{s_0}{n} [U([n]) - U(\emptyset)]$  and  $\mathcal{T}_{\text{sum}} = \sum_{s=1}^{n-1} \mathcal{T}(s)$ . The first term  $G_0$  requires only two utility evaluations. We thus focus on the computational complexity in approximating the second term  $\mathcal{T}_{\text{sum}}$ . To analyze this term, we plug the theorem's assumption that  $\beta(s) = O(1/s)$  into Lemma 8. Specifically, this assumption implies that  $\sum_{s=\bar{s}}^{n-1} \beta(s) = \sum_{s=\bar{s}}^{n-1} O(1/s) = O(\log n)$ . Therefore, from Lemma 7, we have

$$|\widehat{\text{FGSV}}(S_0) - \text{FGSV}(S_0)| \lesssim \bar{s} \sqrt{\frac{\log(n/\delta)}{m_1}} + \alpha_0(1 - \alpha_0) \sqrt{\frac{\log(n/\delta)}{m_2}} \log n + \alpha_0(1 - \alpha_0) \bar{s}^{-v}.$$

Also, the number of utility evaluations, as in the displayed formula in Lemma 8, becomes:

$$N_{\text{eval}} = O \left( \epsilon^{-\frac{4+2v}{v}} \log(n/\delta) + n \left[ 1 + \epsilon^{-2} (\alpha_0(1 - \alpha_0))^2 (\log n)^2 \log(n/\delta) \right] \right).$$

While treating  $\epsilon, \delta$ , and  $v$  as constants, the second term, which is  $O(n [1 + (\alpha_0(1 - \alpha_0))^2 (\log n)^3])$ , dominates the first term (which has a lower power of  $n$ ). Thus, under the condition  $\beta(s) = O(1/s)$ , the total number of utility evaluations simplifies to  $O(n [1 + (\alpha_0(1 - \alpha_0))^2 (\log n)^3])$ .  $\square$

### A.5.1 Technical lemmas used in the proof of Theorem 3

**Lemma 7.** (*Error Bound for  $\hat{\mathcal{T}}_{\text{sum}}$* ) Let  $\mathcal{T}_{\text{sum}} = \sum_{s=1}^{n-1} \mathcal{T}(s)$  and  $\hat{\mathcal{T}}_{\text{sum}} = \sum_{s=1}^{n-1} \hat{\mathcal{T}}(s)$  be the estimate computed by Algorithm 1 using threshold  $\bar{s}$  and sample sizes  $m_1, m_2$ . Under Assumptions 1 and 2, and assuming that the utility function  $U$  is  $\beta(s)$ -deletion stable, for any  $\delta \in (0, 1)$ , with probability at least  $1 - \delta$ :

$$|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \lesssim \bar{s} \sqrt{\frac{\log(n/\delta)}{m_1}} + \alpha_0(1 - \alpha_0) \sqrt{\frac{\log(n/\delta)}{m_2}} \sum_{s=\bar{s}}^{n-1} \beta(s) + \alpha_0(1 - \alpha_0) O(\bar{s}^{-v}).$$

*Proof of Lemma 7.* The total approximation error is bounded using the triangle inequality:

$$|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \leq \sum_{s=1}^{n-1} |\hat{\mathcal{T}}(s) - \mathcal{T}(s)| = \underbrace{\sum_{s=1}^{\bar{s}-1} |\hat{\mathcal{T}}(s) - \mathcal{T}(s)|}_{E_{\text{small}}} + \underbrace{\sum_{s=\bar{s}}^{n-1} |\hat{\mathcal{T}}(s) - \mathcal{T}(s)|}_{E_{\text{large}}}.$$

We first establish concentration bounds for our Monte Carlo estimators. Under Assumption 1,  $|U(S)| \leq C$ . For  $\hat{\mu}_{m_1}$  (Eq. (10)), which averages  $m_1$  terms  $U(S^{(j)})$  bounded in  $[-C, C]$  (range  $2C$ ), Hoeffding's inequality implies that for any specific  $\mu(\frac{s_1}{s}; s, s_0, n)$  and any  $\delta_1 > 0$ , with probability at least  $1 - \delta_1$ :

$$\left| \hat{\mu}_{m_1} \left( \frac{s_1}{s}; s, s_0, n \right) - \mu \left( \frac{s_1}{s}; s, s_0, n \right) \right| \leq C \sqrt{\frac{2 \log(2/\delta_1)}{m_1}} \asymp \sqrt{\frac{\log(1/\delta_1)}{m_1}}. \quad (14)$$

For  $\widehat{\Delta\mu}_{m_2}$  (Eq. (11)), it averages terms  $U(S^{(j)} \cup \{i_1^{(j)}\}) - U(S^{(j)} \cup \{i_2^{(j)}\})$ . By Definition 4, these terms are bounded in  $[-\beta(s), \beta(s)]$ , for any specific  $\Delta\mu(\frac{s_1^*}{s}; s, s_0, n)$  and  $\delta_2 > 0$ , with probability at least  $1 - \delta_2$ :

$$\left| \widehat{\Delta\mu}_{m_2} \left( \frac{s_1^*}{s}; s, s_0, n \right) - \Delta\mu \left( \frac{s_1^*}{s}; s, s_0, n \right) \right| \lesssim \beta(s) \sqrt{\frac{\log(1/\delta_2)}{m_2}}. \quad (15)$$

**Bounding  $E_{\text{small}}$  (Error for  $s < \bar{s}$ ):** For each  $s < \bar{s}$ , the estimate  $\widehat{\mathcal{T}}(s)$  is computed using  $\widehat{\mu}_{m_1}$  within the definition of  $\mathcal{T}(s)$  in (8). This definition involves a sum over approximately  $O(s)$  different values of  $s_1$ , each requiring an estimate  $\widehat{\mu}_{m_1}(\frac{s_1}{s}; s, s_0, n)$ . The coefficient multiplying each  $\mu(\frac{s_1}{s}; s, s_0, n)$  term ( $\text{coeff}(s, s_1) = \frac{n}{s(n-s)}(s_1 - s\alpha_0)$ ) is  $O(1)$ , as established by analyzing its equivalent forms:

$$\text{Form 1: } \frac{n}{n-s} \left( \frac{s_1}{s} - \alpha_0 \right) \quad \text{and} \quad \text{Form 2: } \frac{n}{s} \left( \alpha_0 - \frac{s_0 - s_1}{n-s} \right).$$

When  $s \leq n/2$ , Form 1 is bounded by  $2 \cdot 1 = 2$ , since  $n/(n-s) \leq 2$  and  $|\frac{s_1}{s} - \alpha_0| \leq 1$ . When  $s > n/2$ , Form 2 is bounded by  $2 \cdot 1 = 2$ , since  $n/s < 2$  and  $|\alpha_0 - \frac{s_0 - s_1}{n-s}|$  can also be shown to be at most 1 under the valid range of  $s_1$ . Thus, the coefficient is  $O(1)$ .

The error  $|\widehat{\mathcal{T}}(s) - \mathcal{T}(s)|$  for a given  $s$  arises from propagating the errors from the  $O(s)$  individual  $\widehat{\mu}_{m_1}$  estimations. The total number of distinct  $\mu(\frac{s_1}{s}; s, s_0, n)$  terms that need to be estimated across all  $s < \bar{s}$  is  $N_{\text{small\_ests}} = \sum_{s=1}^{\bar{s}-1} O(s) = O(\bar{s}^2)$ . To ensure all these  $N_{\text{small\_ests}}$  estimations are accurate simultaneously with high probability, we apply a union bound. We set the failure probability for each individual  $\widehat{\mu}_{m_1}$  estimation in (14) to  $\delta_1 = \delta/(2N_{\text{small\_ests}})$ . Consequently, with probability at least  $1 - \delta/2$ , each  $|\widehat{\mu}_{m_1} - \mu| \lesssim \sqrt{\log(N_{\text{small\_ests}}/\delta)/m_1}$ . The error for a single  $\widehat{\mathcal{T}}(s)$  is bounded by  $\sum_{s_1} \mathbb{P}(s_1) |\text{coeff}(s, s_1)| |\widehat{\mu}_{m_1} - \mu| \lesssim O(1) \sqrt{\log(N_{\text{small\_ests}}/\delta)/m_1}$ , since  $\sum_{s_1} \mathbb{P}(s_1) = 1$  and coefficients are  $O(1)$ . Summing these errors over  $s = 1, \dots, \bar{s} - 1$ , we yield that with probability  $1 - \delta/2$ ,

$$E_{\text{small}} = \sum_{s=1}^{\bar{s}-1} |\widehat{\mathcal{T}}(s) - \mathcal{T}(s)| \lesssim \sum_{s=1}^{\bar{s}-1} O(1) \sqrt{\frac{\log(N_{\text{small\_ests}}/\delta)}{m_1}} \lesssim \bar{s} \sqrt{\frac{\log(\bar{s}^2/\delta)}{m_1}}.$$

For conciseness in the final error bound of the lemma, we replace  $\log(\bar{s}^2/\delta)$  with the potentially looser but simpler  $\log(n/\delta)$ , yielding:  $E_{\text{small}} \lesssim \bar{s} \sqrt{\log(n/\delta)/m_1}$ .

**Bounding  $E_{\text{large}}$  (Error for  $s \geq \bar{s}$ ):** For  $s \geq \bar{s}$ ,  $\widehat{\mathcal{T}}(s)$  is computed using the approximation from Proposition 2 (Eq. (9)) with  $\widehat{\Delta\mu}_{m_2}$ . The error has two parts:

$$\begin{aligned} |\widehat{\mathcal{T}}(s) - \mathcal{T}(s)| &\leq \left| \frac{n\alpha_0(1-\alpha_0)}{n-1} \left( \widehat{\Delta\mu}_{m_2} \left( \frac{s_1^*}{s}; s, s_0, n \right) - \Delta\mu \left( \frac{s_1^*}{s}; s, s_0, n \right) \right) \right| \\ &\quad + \left| \frac{n\alpha_0(1-\alpha_0)}{n-1} O(s^{-(1+\nu)}) \right| \\ &\lesssim \alpha_0(1-\alpha_0) \left| \widehat{\Delta\mu}_{m_2} \left( \frac{s_1^*}{s}; s, s_0, n \right) - \Delta\mu \left( \frac{s_1^*}{s}; s, s_0, n \right) \right| + \alpha_0(1-\alpha_0) O(s^{-(1+\nu)}). \end{aligned}$$

Let  $N_{\text{large\_ests}} = n - \bar{s} + 1 \approx n$ . Using a union bound for the  $\widehat{\Delta\mu}_{m_2}$  estimates (setting  $\delta_2 = \delta/(2N_{\text{large\_ests}})$ ), with probability at least  $1 - \delta/2$ :

$$\begin{aligned} E_{\text{large}} &= \sum_{s=\bar{s}}^{n-1} |\widehat{\mathcal{T}}(s) - \mathcal{T}(s)| \\ &\lesssim \alpha_0(1-\alpha_0) \sum_{s=\bar{s}}^{n-1} \beta(s) \sqrt{\frac{\log(N_{\text{large\_ests}}/\delta)}{m_2}} + \alpha_0(1-\alpha_0) \sum_{s=\bar{s}}^{n-1} O(s^{-(1+\nu)}) \\ &\lesssim \alpha_0(1-\alpha_0) \sqrt{\frac{\log(n/\delta)}{m_2}} \sum_{s=\bar{s}}^{n-1} \beta(s) + \alpha_0(1-\alpha_0) O(\bar{s}^{-\nu}), \end{aligned}$$

where the last inequality uses the fact  $\sum_{s=\bar{s}}^{n-1} s^{-(1+v)} = O(\bar{s}^{-v})$  for  $v > 0$ .

**Total Error:** Combining  $E_{\text{small}}$  and  $E_{\text{large}}$ , with overall probability at least  $1 - \delta$  (by union bound on the two failure probabilities  $\delta/2$ ):

$$|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \lesssim \bar{s} \sqrt{\frac{\log(n/\delta)}{m_1}} + \alpha_0(1 - \alpha_0) \sqrt{\frac{\log(n/\delta)}{m_2}} \sum_{s=\bar{s}}^{n-1} \beta(s) + \alpha_0(1 - \alpha_0) O(\bar{s}^{-v}).$$

This completes the proof.  $\square$

**Lemma 8.** (Computational Cost for  $(\epsilon, \delta)$ -Approximation of  $\mathcal{T}_{\text{sum}}$ ) To achieve an  $(\epsilon, \delta)$ -approximation of  $\mathcal{T}_{\text{sum}} = \sum_{s=1}^{n-1} \mathcal{T}(s)$  (i.e., ensuring  $|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \leq \epsilon$  with probability at least  $1 - \delta$ ) using Algorithm 1, with parameters  $m_1, m_2, \bar{s}$  chosen optimally based on Lemma 7, the total number of utility function evaluations is:

$$O \left( \epsilon^{-\frac{4+2v}{v}} \log(n/\delta) + n \left[ 1 + \epsilon^{-2}(\alpha_0(1 - \alpha_0))^2 \left( \sum_{s=\bar{s}}^{n-1} \beta(s) \right)^2 \log(n/\delta) \right] \right).$$

*Proof.* The total number of utility evaluations  $N_{\text{eval}}$  in Algorithm 1 is approximately  $O(\bar{s}^2 m_1 + (n - \bar{s}) m_2)$ , which can be simplified to  $O(\bar{s}^2 m_1 + n m_2)$  as  $\bar{s} \leq n$ .

To achieve the target error  $|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \leq \epsilon$  with probability at least  $1 - \delta$ , we refer to the error bound in Lemma 7:

$$|\hat{\mathcal{T}}_{\text{sum}} - \mathcal{T}_{\text{sum}}| \lesssim \underbrace{\bar{s} \sqrt{\frac{\log(n/\delta)}{m_1}}}_{\text{Term A}} + \underbrace{\alpha_0(1 - \alpha_0) \sqrt{\frac{\log(n/\delta)}{m_2}} \sum_{s=\bar{s}}^{n-1} \beta(s)}_{\text{Term B}} + \underbrace{\alpha_0(1 - \alpha_0) O(\bar{s}^{-v})}_{\text{Term C}}.$$

We set parameters such that each dominant error component contributing to the bound in Lemma 7 is  $O(\epsilon)$ . This involves balancing Term A, Term B, and Term C. The choices for  $m_1, m_2$ , and  $\bar{s}$  that achieve this balance and ensure the total error is  $O(\epsilon)$  are:

$$\begin{aligned} \bar{s} &\asymp \epsilon^{-\frac{1}{v}}, \\ m_1 &\asymp \epsilon^{-\frac{2+2v}{v}} \log(n/\delta), \\ m_2 &\asymp \max \left\{ 1, \epsilon^{-2}(\alpha_0(1 - \alpha_0))^2 \left( \sum_{s=\bar{s}}^{n-1} \beta(s) \right)^2 \log(n/\delta) \right\}. \end{aligned}$$

Substituting these choices of  $m_1, m_2$ , and  $\bar{s}$  into the expression for  $N_{\text{eval}} = O(\bar{s}^2 m_1 + n m_2)$ , we yield

$$\begin{aligned} N_{\text{eval}} &= O \left( \epsilon^{-\frac{4+2v}{v}} \log(n/\delta) + \max \left\{ n \epsilon^{-2}(\alpha_0(1 - \alpha_0))^2 \left( \sum_{s=\bar{s}}^{n-1} \beta(s) \right)^2 \log(n/\delta), n \right\} \right) \\ &= O \left( \epsilon^{-\frac{4+2v}{v}} \log(n/\delta) + n \left[ 1 + \epsilon^{-2}(\alpha_0(1 - \alpha_0))^2 \left( \sum_{s=\bar{s}}^{n-1} \beta(s) \right)^2 \log(n/\delta) \right] \right), \end{aligned}$$

which completes the proof.  $\square$

## A.6 Proof of Proposition 2

Let  $\mathcal{S} = \{z_i^{\mathcal{S}}\}_{i=1}^s \subset \mathcal{Z}^s$  denote the base training set, and let  $z_1, z'_1, z_2, z'_2 \in \mathcal{Z}$  be four additional data points. We construct four augmented training data sequences, where the  $i$ -th element of a sequence

$z^{(j,k)}$  is denoted  $z_i^{(j,k)}$  and defined as follows:

$$\begin{aligned} z_i^{(1,1)} &= \begin{cases} z_i^S, & 1 \leq i \leq s, \\ z_1, & i = s+1, \\ z'_1, & i = s+2, \end{cases} & z_i^{(1,2)} &= \begin{cases} z_i^S, & 1 \leq i \leq s, \\ z_1, & i = s+1, \\ z_2, & i = s+2, \end{cases} \\ z_i^{(2,1)} &= \begin{cases} z_i^S, & 1 \leq i \leq s, \\ z'_2, & i = s+1, \\ z'_1, & i = s+2, \end{cases} & z_i^{(2,2)} &= \begin{cases} z_i^S, & 1 \leq i \leq s, \\ z'_2, & i = s+1, \\ z_2, & i = s+2. \end{cases} \end{aligned}$$

By construction, the first  $s$  entries (the base set  $\mathcal{S}$ ) are shared across all four sequences. The differences between the sequences are confined to the  $(s+1)$ -th and  $(s+2)$ -th positions. Specifically:

- For the  $(s+1)$ -th position:  $z_{s+1}^{(1,1)} = z_{s+1}^{(1,2)} (= z_1)$ , and  $z_{s+1}^{(2,1)} = z_{s+1}^{(2,2)} (= z'_2)$ .
- For the  $(s+2)$ -th position:  $z_{s+2}^{(1,1)} = z_{s+2}^{(2,1)} (= z'_1)$ , and  $z_{s+2}^{(1,2)} = z_{s+2}^{(2,2)} (= z_2)$ .

We introduce the notation for the SGD trajectories **under the same random seed**. Fixing a random seed implies a shared sequence of mini-batch index sets  $(I_1, \dots, I_T)$  for training. The corresponding SGD trajectories are denoted  $w_t^{(j,k)}$  ( $j, k \in \{1, 2\}$ ), representing the model parameters after  $t$  SGD steps. Specifically, they can be represented as:

$$\begin{aligned} w_t^{(1,1)} &= w_t \left( z_{I_1}^{(1,1)}, \dots, z_{I_t}^{(1,1)} \right), \\ w_t^{(1,2)} &= w_t \left( z_{I_1}^{(1,2)}, \dots, z_{I_t}^{(1,2)} \right), \\ w_t^{(2,1)} &= w_t \left( z_{I_1}^{(2,1)}, \dots, z_{I_t}^{(2,1)} \right), \\ w_t^{(2,2)} &= w_t \left( z_{I_1}^{(2,2)}, \dots, z_{I_t}^{(2,2)} \right). \end{aligned} \tag{16}$$

Here,  $w_t(\cdot)$  is a function that takes a sequence of mini-batch data points as inputs and outputs a parameter vector;  $z_I^{(j,k)} = \{z_i^{(j,k)} : i \in I\}$  denotes a mini-batch constructed from the data sequence  $z^{(j,k)}$  using the index set  $I$ . The arguments  $z_{I_\tau}^{(j,k)}$  in Eq. (16) thus represent the specific mini-batches used at each step  $\tau \in [1, t]$ . The variations among these trajectories arise only from the different data points at indices  $s+1$  and  $s+2$  within their respective sequences. This setup focuses on analyzing how these specific data perturbations affect the SGD iterations.

In this section, for clarity in tracking the dependence on the number of training steps  $T$ , we use  $U(\cdot; T)$  to denote the expected utility after  $T$  SGD steps, as opposed to  $U(\cdot)$  used in the main text.

To demonstrate the second-order stability of the utility function  $U(\cdot; T)$ , as stated in Proposition 2, our approach relies on key intermediate results that connect utility differences to parameter stability. Lemma 9 below establishes this crucial link.

**Lemma 9.** *Let  $\mathcal{S} \in \mathcal{Z}^s$  be a base dataset, and  $z_1, z'_1, z_2, z'_2 \in \mathcal{Z}$  be additional data points. Recall that  $U(\mathcal{X}; T)$  denotes the expected utility of a model trained on dataset  $\mathcal{X}$  for  $T$  SGD iterations. Then, under Assumption 3, we have:*

$$\begin{aligned} &|U(\mathcal{S} \cup \{z_1, z'_1\}; T) - U(\mathcal{S} \cup \{z_1, z_2\}; T) - U(\mathcal{S} \cup \{z'_1, z'_2\}; T) + U(\mathcal{S} \cup \{z_2, z'_2\}; T)| \\ &\lesssim \mathbb{E}_{I_1, \dots, I_T} \left\| w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)} \right\| + \max_{j_1, k_1, j_2, k_2 \in \{1, 2\}} \mathbb{E}_{I_1, \dots, I_T} \left\| w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)} \right\|^2, \end{aligned}$$

where  $w_T^{(j,k)}$  denotes the final SGD iterate trained using the shared mini-batch sequence  $(I_1, \dots, I_T)$  on the data sequence  $z^{(j,k)}$  defined in Eq. (16).

**Remark 1.** *The utility terms  $U(\mathcal{S} \cup \{\cdot, \cdot\}; T)$  on the left-hand side of the inequality in Lemma 9 are, by definition, expectations over independent SGD runs (i.e., each utility term averages over its own random mini-batch selections). In contrast, the parameter differences on the right-hand side are for iterates  $w_T^{(j,k)}$  that are explicitly trained using a common, shared sequence of mini-batch draws  $(I_1, \dots, I_T)$ . This formulation is key, as it leverages the shared randomness for a tighter analysis of parameter stability.*

Having established the connection between the utility's stability and the SGD iterates via Lemma 9, the next crucial step is to quantify the magnitude of these iterate differences. Specifically, the two terms on the right-hand side of the inequality in Lemma 9—the expected  $L_2$ -norm of the second-order parameter difference,  $\mathbb{E}_{I_1, \dots, I_T} \|w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)}\|$ , and the maximum expected squared  $L_2$ -norm of pairwise parameter differences,  $\max \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\|^2$ —need to be controlled. The following two lemmas provide the necessary explicit upper bounds for these quantities.

**Lemma 10.** *Under Assumption 3, and choosing step size  $\alpha_t \leq \frac{c}{t}$ , we have:*

$$\mathbb{E}_{I_1, \dots, I_t} \|w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)}\| \leq \frac{24\rho L^2}{(s+2)\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t^{2c\beta} + \frac{8cL}{(s+2)^2} t^{c\beta} \log t.$$

**Lemma 11.** *Under Assumption 3, if the learning rate satisfies  $\alpha_t \leq \frac{c}{t}$ , then for any  $j_1, k_1, j_2, k_2 \in \{1, 2\}$ :*

$$\mathbb{E}_{I_1, \dots, I_t} \|w_t^{(j_1, k_1)} - w_t^{(j_2, k_2)}\|^2 \leq \frac{16L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right).$$

Equipped with these results, we are ready to prove Proposition 2.

*Proof of Proposition 2.* By Lemma 10 and Lemma 11, we have

$$\begin{aligned} \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)}\| &\lesssim \frac{T^{2c\beta}}{s^2} + \frac{c^2 T^{2c\beta}}{sm} + \frac{c T^{c\beta} \log T}{s^2}, \\ \max_{j_1, k_1, j_2, k_2} \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\| &\lesssim \frac{T^{2c\beta}}{s^2} + \frac{c^2 T^{2c\beta}}{sm} \end{aligned}$$

Using the substitutions  $c \asymp s^{-\tau_1}$ ,  $m \asymp s^{\tau_2}$ , and  $T \asymp s^{\tau_3}$ , we compute

$$\begin{aligned} \frac{T^{2c\beta}}{s^2} &\asymp \frac{1}{s^2} \cdot \exp(2c\beta \log T) = \frac{1}{s^2} \cdot \exp(2\beta s^{-\tau_1} \cdot \tau_3 \log s), \\ \frac{c^2 T^{2c\beta}}{sm} &\asymp \frac{1}{s^{1+2\tau_1+\tau_2}} \cdot \exp(2\beta s^{-\tau_1} \cdot \tau_3 \log s), \\ \frac{c \log T \cdot T^{c\beta}}{s^2} &\asymp \frac{s^{-\tau_1} \cdot \tau_3 \log s}{s^2} \cdot \exp(\beta s^{-\tau_1} \cdot \tau_3 \log s). \end{aligned}$$

Using the fact that  $\exp(\gamma s^{-\tau_1} \log s) = 1 + o(1)$  for any  $\gamma, \tau_1 > 0$ , we conclude:

$$\begin{aligned} \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)}\| &\lesssim \frac{1}{s^2} + \frac{1}{s^{1+2\tau_1+\tau_2}}, \\ \max_{j_1, k_1, j_2, k_2} \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\| &\lesssim \frac{1}{s^2} + \frac{1}{s^{1+2\tau_1+\tau_2}}. \end{aligned}$$

Since  $v = 2\tau_1 + \tau_2 - \frac{1}{2}$ , then

$$\begin{aligned} \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)}\| &\lesssim \frac{1}{s^{\frac{3}{2} + \min\{\frac{1}{2}, v\}}}, \\ \max_{j_1, k_1, j_2, k_2} \mathbb{E}_{I_1, \dots, I_T} \|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\| &\lesssim \frac{1}{s^{\frac{3}{2} + \min\{\frac{1}{2}, v\}}}. \end{aligned}$$

Finally, applying Lemma 9 yields the desired second-order stability bound.  $\square$

### A.6.1 Preliminary first-order stability results and proof of Lemma 11

To build towards the proof of Lemma 10, we first analyze first-order algorithmic stability. This serves to introduce key proof techniques in a simplified setting and provides intermediate results that are essential for the subsequent second-order analysis. The ideas presented here are closely aligned with those in Hardt et al. [5].

Consider a base dataset  $\mathcal{S} = \{z_i^{\mathcal{S}}\}_{i=1}^s \subset \mathcal{Z}^s$ . We introduce two additional, distinct data points  $z_a, z_b \in \mathcal{Z}$ . We then define two augmented data sequences,  $z^{(a)}$  and  $z^{(b)}$ , each of length  $s + 1$ :

$$z_i^{(a)} = \begin{cases} z_i^{\mathcal{S}}, & 1 \leq i \leq s, \\ z_a, & i = s + 1, \end{cases} \quad z_i^{(b)} = \begin{cases} z_i^{\mathcal{S}}, & 1 \leq i \leq s, \\ z_b, & i = s + 1. \end{cases}$$

These sequences  $z^{(a)}$  and  $z^{(b)}$  differ only in their  $(s + 1)$ -th element. Let  $w_t^{(a)}$  and  $w_t^{(b)}$  denote the iterates produced by SGD after  $t$  steps when trained on their respective data sequences, using a shared sequence of mini-batch indices  $(I_1, \dots, I_t)$ . Consistent with Eq. (16), these are:

$$w_t^{(a)} = w_t(z_{I_1}^{(a)}, \dots, z_{I_t}^{(a)}), \\ w_t^{(b)} = w_t(z_{I_1}^{(b)}, \dots, z_{I_t}^{(b)}).$$

We now state two standard first-order stability results concerning such iterates.

**Lemma 12.** *Under Assumption 3, if the learning rate satisfies  $\alpha_t \leq \frac{c}{t}$ , then for iterates  $w_t^{(a)}$  and  $w_t^{(b)}$  (trained on data sequences  $z^{(a)}$  and  $z^{(b)}$  of effective size  $N = s + 1$  that differ by one point):*

$$\mathbb{E}_{I_1, \dots, I_t} \|w_t^{(a)} - w_t^{(b)}\| \leq \frac{2L}{\beta N} t^{c\beta}.$$

**Lemma 13.** *Under Assumption 3, if the learning rate satisfies  $\alpha_t \leq \frac{c}{t}$ , then for iterates  $w_t^{(a)}$  and  $w_t^{(b)}$  (trained on data sequences  $z^{(a)}$  and  $z^{(b)}$  of effective size  $N = s + 1$  that differ by one point):*

$$\mathbb{E}_{I_1, \dots, I_t} \|w_t^{(a)} - w_t^{(b)}\|^2 \leq \frac{4L^2}{N} t^{2c\beta} \left( \frac{1}{N\beta^2} + \frac{2c^2}{m} \right).$$

Note: In the context of these lemmas,  $N$  represents the total number of data points in the sequences being compared. When applying these to the iterates  $w_t^{(j,k)}$  (which are trained on data sequences of length  $s + 2$ ),  $N$  will correspond to  $s + 2$ .

These foundational first-order stability results, particularly Lemma 13, allow us to derive Lemma 11. We restate Lemma 11 for clarity before its proof.

**Lemma 11.** *Under Assumption 3, if the learning rate satisfies  $\alpha_t \leq \frac{c}{t}$ , then for any  $j_1, k_1, j_2, k_2 \in \{1, 2\}$ :*

$$\mathbb{E}_{I_1, \dots, I_t} \|w_t^{(j_1, k_1)} - w_t^{(j_2, k_2)}\|^2 \leq \frac{16L^2}{s + 2} t^{2c\beta} \left( \frac{1}{(s + 2)\beta^2} + \frac{2c^2}{m} \right).$$

*Proof of Lemma 11.* The data sequences  $z^{(j_1, k_1)}$  and  $z^{(j_2, k_2)}$  (each of size  $N = s + 2$ ) differ at most at two positions (index  $s + 1$  and  $s + 2$ ). We can introduce an intermediate iterate,  $w_t^{(j_1, k_2)}$ . Using the triangle inequality and the property  $\|x + y\|^2 \leq 2(\|x\|^2 + \|y\|^2)$ :

$$\begin{aligned} \mathbb{E}_{I_1, \dots, I_t} \|w_t^{(j_1, k_1)} - w_t^{(j_2, k_2)}\|^2 &= \mathbb{E}_{I_1, \dots, I_t} \|(w_t^{(j_1, k_1)} - w_t^{(j_1, k_2)}) + (w_t^{(j_1, k_2)} - w_t^{(j_2, k_2)})\|^2 \\ &\leq 2 \left\{ \mathbb{E}_{I_1, \dots, I_t} \|w_t^{(j_1, k_1)} - w_t^{(j_1, k_2)}\|^2 + \mathbb{E}_{I_1, \dots, I_t} \|w_t^{(j_1, k_2)} - w_t^{(j_2, k_2)}\|^2 \right\}. \end{aligned}$$

The term  $\|w_t^{(j_1, k_1)} - w_t^{(j_1, k_2)}\|^2$  involves iterates from data sequences  $z^{(j_1, k_1)}$  and  $z^{(j_1, k_2)}$ . These sequences share the same first  $s + 1$  points (namely,  $z_1^{\mathcal{S}}, \dots, z_s^{\mathcal{S}}, z_{s+1}^{(j_1)}$ ) and differ only at the  $(s + 2)$ -th position. This is a one-point difference in data sequences of effective size  $N = s + 2$ . Similarly, the term  $\|w_t^{(j_1, k_2)} - w_t^{(j_2, k_2)}\|^2$  involves iterates from data sequences  $z^{(j_1, k_2)}$  and  $z^{(j_2, k_2)}$ . These sequences share the first  $s$  points ( $z_1^{\mathcal{S}}, \dots, z_s^{\mathcal{S}}$ ) and the  $(s + 2)$ -th point ( $z_{s+2}^{(k_2)}$ ), differing only at the  $(s + 1)$ -th position. This is also a one-point difference in data sequences of effective size  $N = s + 2$ .

Therefore, applying Lemma 13 to each of these one-point difference terms (with  $N = s + 2$  as the effective dataset size) yields:

$$\begin{aligned}\mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(j_1, k_1)} - w_t^{(j_1, k_2)} \right\|^2 &\leq \frac{4L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right), \\ \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(j_1, k_2)} - w_t^{(j_2, k_2)} \right\|^2 &\leq \frac{4L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right).\end{aligned}$$

Substituting these into the inequality:

$$\begin{aligned}\mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(j_1, k_1)} - w_t^{(j_2, k_2)} \right\|^2 &\leq 2 \times 2 \times \left\{ \frac{8L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) \right\} \\ &= \frac{16L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right).\end{aligned}$$

This completes the proof.  $\square$

We now turn to the second-order behavior of stochastic gradient methods. Analyzing this provides a more refined understanding of how small perturbations to the training dataset influence the final learned parameters. The analysis relies on the higher-order smoothness properties of the loss function  $\ell(w; z)$  as detailed in Assumption 3, particularly the Lipschitz continuity of its Hessian (second derivative with respect to  $w$ ). The following lemma, which is a consequence of this Lipschitz Hessian property, serves as a key technical tool for bounding second-order differences of gradients.

**Lemma 14.** *Under Assumption 3, the loss function  $\ell(w; z)$  satisfies the following second-order approximation bound for all  $w, h \in \mathbb{R}^d$  and  $z \in \mathcal{Z}$ :*

$$\left\| \nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) - \nabla_w^2 \ell(w; z) h \right\| \leq \rho \|h\|^2.$$

With this tool, we now begin the proof of Lemma 10, which bounds the expected  $L_2$ -norm of the second-order difference of the SGD iterates. We restate Lemma 10 for clarity before its proof.

**Lemma 10.** *Under Assumption 3, and choosing step size  $\alpha_t \leq \frac{c}{t}$ , we have:*

$$\mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| \leq \frac{24\rho L^2}{(s+2)\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t^{2c\beta} + \frac{8cL}{(s+2)^2} t^{c\beta} \log t.$$

*Proof of Lemma 10.* From the SGD update rule, the difference between the second-order differences of the iterates at step  $t$  and  $t-1$  can be bounded:

$$\begin{aligned}&\left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| - \left\| w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)} \right\| \\ &\leq \left\| \frac{\alpha_t}{m} \sum_{i \in I_t} \left[ \nabla_w \ell(w_{t-1}^{(1,1)}; z_i^{(1,1)}) - \nabla_w \ell(w_{t-1}^{(1,2)}; z_i^{(1,2)}) - \nabla_w \ell(w_{t-1}^{(2,1)}; z_i^{(2,1)}) + \nabla_w \ell(w_{t-1}^{(2,2)}; z_i^{(2,2)}) \right] \right\|.\end{aligned}$$

Let's define the term inside the norm on the right-hand side, representing the sum of gradient differences for a given mini-batch  $I_t$ :

$$\Delta(t-1, i) := \nabla_w \ell(w_{t-1}^{(1,1)}; z_i^{(1,1)}) - \nabla_w \ell(w_{t-1}^{(1,2)}; z_i^{(1,2)}) - \nabla_w \ell(w_{t-1}^{(2,1)}; z_i^{(2,1)}) + \nabla_w \ell(w_{t-1}^{(2,2)}; z_i^{(2,2)}).$$

Taking the expectation with respect to the mini-batch sequences  $I_1, \dots, I_t$ , we have:

$$\begin{aligned}
& \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| - \mathbb{E}_{I_1, \dots, I_{t-1}} \left\| w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)} \right\| \\
&= \mathbb{E}_{I_1, \dots, I_{t-1}} \left\{ \mathbb{E}_{I_t} \left[ \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| \right] - \left\| w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)} \right\| \right\} \\
&\leq \mathbb{E}_{I_1, \dots, I_{t-1}} \left\{ \mathbb{E}_{I_t} \left[ \left\| \frac{\alpha_t}{m} \sum_{i \in I_t} \Delta(t-1, i) \right\| \right] \right\} \\
&\leq \mathbb{E}_{I_1, \dots, I_{t-1}} \left\{ \frac{\alpha_t}{m} \mathbb{E}_{I_t} \left[ \sum_{i \in I_t} \|\Delta(t-1, i)\| \right] \right\} \\
&= \mathbb{E}_{I_1, \dots, I_{t-1}} \left\{ \frac{\alpha_t}{m} \sum_{i=1}^{s+2} \|\Delta(t-1, i)\| \cdot \mathbb{P}(z_j \in I_t) \right\} \\
&= \frac{\alpha_t}{s+2} \sum_{i=1}^{s+2} \mathbb{E}_{I_1, \dots, I_{t-1}} \|\Delta(t-1, i)\| \tag{17}
\end{aligned}$$

In the last step,  $\mathbb{P}(i \in I_t) = \frac{m}{s+2}$  because  $I_t$  is sampled uniformly from the  $s+2$  points.

Now, we analyze the bounds for  $\mathbb{E}_{I_1, \dots, I_{t-1}} \|\Delta(t-1, i)\|$ .

**Case 1:  $i \leq s$ .** For these points,  $z_i^{(1,1)} = z_i^{(1,2)} = z_i^{(2,1)} = z_i^{(2,2)} = z_i^S$ . Applying Lemma 14 (which relies on the Lipschitz Hessian property):

$$\begin{aligned}
\|\Delta(t-1, i)\| &\leq \left\| \nabla_w^2 \ell(w_{t-1}^{(1,1)}; z_i^S) \left[ -(w_{t-1}^{(1,2)} - w_{t-1}^{(1,1)}) - (w_{t-1}^{(2,1)} - w_{t-1}^{(1,1)}) + (w_{t-1}^{(2,2)} - w_{t-1}^{(1,1)}) \right] \right\| \\
&\quad + \rho \|w_{t-1}^{(1,2)} - w_{t-1}^{(1,1)}\|^2 + \rho \|w_{t-1}^{(2,1)} - w_{t-1}^{(1,1)}\|^2 + \rho \|w_{t-1}^{(2,2)} - w_{t-1}^{(1,1)}\|^2 \\
&\leq \beta \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)}\| \\
&\quad + \rho \|w_{t-1}^{(1,2)} - w_{t-1}^{(1,1)}\|^2 + \rho \|w_{t-1}^{(2,1)} - w_{t-1}^{(1,1)}\|^2 \\
&\quad + 2\rho \left( \|w_{t-1}^{(2,2)} - w_{t-1}^{(1,2)}\|^2 + \|w_{t-1}^{(1,2)} - w_{t-1}^{(1,1)}\|^2 \right).
\end{aligned}$$

Taking the expectation and applying Lemma 13 for the squared difference terms:

$$\begin{aligned}
\mathbb{E}_{I_1, \dots, I_{t-1}} \|\Delta(t-1, i)\| &\leq \beta \cdot \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)}\| \\
&\quad + \rho \cdot \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,2)} - w_{t-1}^{(1,1)}\|^2 + \rho \cdot \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(2,1)} - w_{t-1}^{(1,1)}\|^2 \\
&\quad + 2\rho \left( \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(2,2)} - w_{t-1}^{(1,2)}\|^2 + \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,2)} - w_{t-1}^{(1,1)}\|^2 \right) \\
&\leq \beta \cdot \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)}\| \\
&\quad + 6\rho \cdot \frac{4L^2}{s+2} (t-1)^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right).
\end{aligned}$$

**Case 2:  $i = s+1$ .** Here,  $z_{s+1}^{(1,1)} = z_{s+1}^{(1,2)} (= z_1)$  and  $z_{s+1}^{(2,1)} = z_{s+1}^{(2,2)} (= z'_2)$ .

$$\begin{aligned}
\Delta(t-1, s+1) &= \left\| (\nabla_w \ell(w_{t-1}^{(1,1)}; z_1) - \nabla_w \ell(w_{t-1}^{(1,2)}; z_1)) - (\nabla_w \ell(w_{t-1}^{(2,1)}; z'_2) - \nabla_w \ell(w_{t-1}^{(2,2)}; z'_2)) \right\| \\
&\leq \left\| \nabla_w \ell(w_{t-1}^{(1,1)}; z_1) - \nabla_w \ell(w_{t-1}^{(1,2)}; z_1) \right\| + \left\| \nabla_w \ell(w_{t-1}^{(2,1)}; z'_2) - \nabla_w \ell(w_{t-1}^{(2,2)}; z'_2) \right\| \\
&\leq \beta \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)}\| + \beta \|w_{t-1}^{(2,1)} - w_{t-1}^{(2,2)}\|.
\end{aligned}$$

Taking expectation and using Lemma 12 (first-order stability for one-point difference):

$$\begin{aligned}
\mathbb{E}_{I_1, \dots, I_{t-1}} \{\Delta(t-1, s+1)\} &\leq \beta \cdot \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)}\| + \beta \cdot \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(2,1)} - w_{t-1}^{(2,2)}\| \\
&\leq \beta \cdot \frac{2L}{\beta(s+2)} (t-1)^{c\beta} + \beta \cdot \frac{2L}{\beta(s+2)} (t-1)^{c\beta} \\
&\leq \frac{4L}{s+2} t^{c\beta}.
\end{aligned}$$



A similar bound holds for  $i = s + 2$ .

Substituting these bounds into Eq. (17): The sum over  $i = 1, \dots, s + 2$  has  $s$  terms from Case 1 and 2 terms from Case 2.

$$\begin{aligned}
& \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| - \mathbb{E}_{I_1, \dots, I_{t-1}} \left\| w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)} \right\| \\
& \leq \frac{\alpha_t s}{s+2} \left[ \beta \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - \dots + w_{t-1}^{(2,2)}\| + \frac{24\rho L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) \right] \\
& \quad + \frac{\alpha_t \cdot 2}{s+2} \left[ \frac{4L}{s+2} t^{c\beta} \right] \\
& \leq \alpha_t \beta \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)}\| \quad (\text{since } s/(s+2) < 1) \\
& \quad + \alpha_t \frac{24\rho L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) + \alpha_t \frac{8L}{(s+2)^2} t^{c\beta}.
\end{aligned}$$

Using  $\alpha_t \leq c/t$ :

$$\begin{aligned}
& \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| \\
& \leq \left( 1 + \frac{c\beta}{t} \right) \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)}\| \\
& \quad + \frac{c}{t} \left[ \frac{24\rho L^2}{s+2} t^{2c\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) + \frac{8L}{(s+2)^2} t^{c\beta} \right] \\
& \leq \exp \left( \frac{c\beta}{t} \right) \mathbb{E}_{I_1, \dots, I_{t-1}} \|w_{t-1}^{(1,1)} - w_{t-1}^{(1,2)} - w_{t-1}^{(2,1)} + w_{t-1}^{(2,2)}\| \\
& \quad + \frac{24c\rho L^2}{s+2} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t^{2c\beta-1} + \frac{8cL}{(s+2)^2} t^{c\beta-1}.
\end{aligned}$$

Unrolling this recurrence from  $t_0 = 1$  to  $t$ :

$$\begin{aligned}
& \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| \\
& \leq \sum_{t_0=1}^t \left[ \frac{24c\rho L^2}{s+2} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t_0^{2c\beta-1} + \frac{8cL}{(s+2)^2} t_0^{c\beta-1} \right] \prod_{k=t_0+1}^t \exp \left( \frac{c\beta}{k} \right) \\
& = \sum_{t_0=1}^t \left[ \frac{24c\rho L^2}{s+2} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t_0^{2c\beta-1} + \frac{8cL}{(s+2)^2} t_0^{c\beta-1} \right] \exp \left( \sum_{k=t_0+1}^t \frac{c\beta}{k} \right).
\end{aligned}$$

Using the approximation  $\sum_{k=t_0+1}^t \frac{1}{k} \approx \log(t/t_0)$ :

$$\begin{aligned}
& \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| \\
& \leq \sum_{t_0=1}^t \left[ \frac{24c\rho L^2}{s+2} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t_0^{2c\beta-1} + \frac{8cL}{(s+2)^2} t_0^{c\beta-1} \right] \left( \frac{t}{t_0} \right)^{c\beta} \\
& = \frac{24c\rho L^2 t^{c\beta}}{s+2} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) \sum_{t_0=1}^t t_0^{c\beta-1} + \frac{8cL t^{c\beta}}{(s+2)^2} \sum_{t_0=1}^t t_0^{-1}.
\end{aligned}$$

Using  $\sum_{t_0=1}^t t_0^{c\beta-1} \approx \frac{t^{c\beta}}{c\beta}$  for  $c\beta > 0$ , and  $\sum_{t_0=1}^t t_0^{-1} \approx \log t$ :

$$\begin{aligned}
& \mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(1,1)} - w_t^{(1,2)} - w_t^{(2,1)} + w_t^{(2,2)} \right\| \\
& \leq \frac{24c\rho L^2 t^{c\beta}}{s+2} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) \frac{t^{c\beta}}{c\beta} + \frac{8cL t^{c\beta}}{(s+2)^2} \log t \\
& = \frac{24\rho L^2}{(s+2)\beta} \left( \frac{1}{(s+2)\beta^2} + \frac{2c^2}{m} \right) t^{2c\beta} + \frac{8cL}{(s+2)^2} t^{c\beta} \log t.
\end{aligned}$$

This completes the proof.  $\square$

### A.6.2 Technical lemmas used in the proof of Proposition 2

**Lemma 9.** Let  $\mathcal{S} \in \mathcal{Z}^s$  be a base dataset, and  $z_1, z'_1, z_2, z'_2 \in \mathcal{Z}$  be additional data points. Recall that  $U(\mathcal{X}; T)$  denotes the expected utility of a model trained on dataset  $\mathcal{X}$  for  $T$  SGD iterations. Then, under Assumption 3, we have:

$$\begin{aligned} & |U(\mathcal{S} \cup \{z_1, z'_1\}; T) - U(\mathcal{S} \cup \{z_1, z_2\}; T) - U(\mathcal{S} \cup \{z'_1, z'_2\}; T) + U(\mathcal{S} \cup \{z_2, z'_2\}; T)| \\ & \lesssim \mathbb{E}_{I_1, \dots, I_T} \left\| w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)} \right\| + \max_{j_1, k_1, j_2, k_2 \in \{1,2\}} \mathbb{E}_{I_1, \dots, I_T} \left\| w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)} \right\|^2, \end{aligned}$$

where  $w_T^{(j,k)}$  denotes the final SGD iterate trained using the shared mini-batch sequence  $(I_1, \dots, I_T)$  on the data sequence  $z^{(j,k)}$  defined in Eq. (16).

*Proof of Lemma 9.* Let  $\mathcal{S}$  be the base dataset. We expand the expression using the definition of  $U(\mathcal{S}; T) = \mathbb{E}[u(w_T)]$  where the expectation is over the choice of mini-batches  $I_1, \dots, I_T$ :

$$\begin{aligned} & |U(\mathcal{S} \cup \{z_1, z'_1\}; T) - U(\mathcal{S} \cup \{z_1, z_2\}; T) - U(\mathcal{S} \cup \{z'_1, z'_2\}; T) + U(\mathcal{S} \cup \{z_2, z'_2\}; T)| \\ & = \left| \mathbb{E}_{I_1, \dots, I_T} \left[ u(w_T^{(1,1)}) \right] - \mathbb{E}_{I_1, \dots, I_T} \left[ u(w_T^{(1,2)}) \right] - \mathbb{E}_{I_1, \dots, I_T} \left[ u(w_T^{(2,1)}) \right] + \mathbb{E}_{I_1, \dots, I_T} \left[ u(w_T^{(2,2)}) \right] \right| \\ & = \left| \mathbb{E}_{I_1, \dots, I_T} \left[ u(w_T^{(1,1)}) - u(w_T^{(1,2)}) - u(w_T^{(2,1)}) + u(w_T^{(2,2)}) \right] \right| \\ & \leq \mathbb{E}_{I_1, \dots, I_T} \left| u(w_T^{(1,1)}) - u(w_T^{(1,2)}) - u(w_T^{(2,1)}) + u(w_T^{(2,2)}) \right|. \quad (\text{by Jensen's inequality}) \end{aligned}$$

We now expand the differences of  $u$  using a Taylor expansion. For  $u(w_A) - u(w_B)$ , by Taylor's theorem with remainder:  $u(w_A) - u(w_B) = \langle \nabla_\theta u(w_B), w_A - w_B \rangle + R_{A,B}$ , where the remainder  $R_{A,B}$  satisfies  $|R_{A,B}| \leq C_u \|w_A - w_B\|^2$  if  $\nabla_\theta u$  is  $C_u$ -Lipschitz (from Assumption 3, using  $C_u$  for this constant). Thus, this remainder is  $O(\|w_A - w_B\|^2)$ .

So, we have:

$$\begin{aligned} u(w_T^{(1,1)}) - u(w_T^{(1,2)}) &= \langle \nabla_\theta u(w_T^{(1,2)}), w_T^{(1,1)} - w_T^{(1,2)} \rangle + O(\|w_T^{(1,1)} - w_T^{(1,2)}\|^2), \\ u(w_T^{(2,1)}) - u(w_T^{(2,2)}) &= \langle \nabla_\theta u(w_T^{(2,2)}), w_T^{(2,1)} - w_T^{(2,2)} \rangle + O(\|w_T^{(2,1)} - w_T^{(2,2)}\|^2). \end{aligned}$$

Then, the term inside the expectation is:

$$\begin{aligned} & u(w_T^{(1,1)}) - u(w_T^{(1,2)}) - (u(w_T^{(2,1)}) - u(w_T^{(2,2)})) \\ &= \langle \nabla_\theta u(w_T^{(1,2)}), w_T^{(1,1)} - w_T^{(1,2)} \rangle - \langle \nabla_\theta u(w_T^{(2,2)}), w_T^{(2,1)} - w_T^{(2,2)} \rangle \\ & \quad + O(\|w_T^{(1,1)} - w_T^{(1,2)}\|^2) + O(\|w_T^{(2,1)} - w_T^{(2,2)}\|^2). \end{aligned}$$

The sum of the  $O(\cdot)$  terms is  $O(\|w_T^{(1,1)} - w_T^{(1,2)}\|^2 + \|w_T^{(2,1)} - w_T^{(2,2)}\|^2)$ .

We manipulate the inner product terms:

$$\begin{aligned} & \langle \nabla_\theta u(w_T^{(1,2)}), w_T^{(1,1)} - w_T^{(1,2)} \rangle - \langle \nabla_\theta u(w_T^{(2,2)}), w_T^{(2,1)} - w_T^{(2,2)} \rangle \\ &= \langle \nabla_\theta u(w_T^{(1,2)}), w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)} \rangle \\ & \quad + \langle \nabla_\theta u(w_T^{(1,2)}) - \nabla_\theta u(w_T^{(2,2)}), w_T^{(2,1)} - w_T^{(2,2)} \rangle. \end{aligned}$$

For the second term in this sum, using Cauchy-Schwarz and the  $C_u$ -Lipschitz continuity of  $\nabla_\theta u$ :

$$\begin{aligned} & \left| \langle \nabla_\theta u(w_T^{(1,2)}) - \nabla_\theta u(w_T^{(2,2)}), w_T^{(2,1)} - w_T^{(2,2)} \rangle \right| \\ & \leq \left\| \nabla_\theta u(w_T^{(1,2)}) - \nabla_\theta u(w_T^{(2,2)}) \right\| \cdot \left\| w_T^{(2,1)} - w_T^{(2,2)} \right\| \\ & \leq C_u \left\| w_T^{(1,2)} - w_T^{(2,2)} \right\| \cdot \left\| w_T^{(2,1)} - w_T^{(2,2)} \right\|. \end{aligned}$$

Using the inequality  $ab \leq \frac{1}{2}(a^2 + b^2)$ , this is bounded by:

$$\frac{C_u}{2} \left( \|w_T^{(1,2)} - w_T^{(2,2)}\|^2 + \|w_T^{(2,1)} - w_T^{(2,2)}\|^2 \right).$$

This term is also of the order  $O\left(\|w_T^{(1,2)} - w_T^{(2,2)}\|^2 + \|w_T^{(2,1)} - w_T^{(2,2)}\|^2\right)$ .

Combining all terms, the absolute value  $\left|u(w_T^{(1,1)}) - u(w_T^{(1,2)}) - u(w_T^{(2,1)}) + u(w_T^{(2,2)})\right|$  is bounded by:

$$\begin{aligned} & \left| \left\langle \nabla_\theta u(w_T^{(1,2)}), w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)} \right\rangle \right| \\ & + O\left(\|w_T^{(1,1)} - w_T^{(1,2)}\|^2 + \|w_T^{(2,1)} - w_T^{(2,2)}\|^2 + \|w_T^{(1,2)} - w_T^{(2,2)}\|^2\right). \end{aligned}$$

Let  $C_{\|\nabla u\|}$  be the bound on  $\|\nabla_\theta u(\cdot)\|$  from Assumption 3. Then the first term is bounded by:

$$C_{\|\nabla u\|} \|w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)}\|.$$

The sum of  $O(\cdot)$  terms involves various pairwise squared differences. Each such term is of the form  $\|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\|^2$ . The sum is thus  $O\left(\max_{j_1, k_1, j_2, k_2} \|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\|^2\right)$ .

Taking the expectation over  $I_1, \dots, I_T$ :

$$\begin{aligned} & |U(\mathcal{S} \cup \{z_1, z'_1\}; T) - U(\mathcal{S} \cup \{z_1, z_2\}; T) - U(\mathcal{S} \cup \{z'_1, z'_2\}; T) + U(\mathcal{S} \cup \{z_2, z'_2\}; T)| \\ & \leq C_{\|\nabla u\|} \mathbb{E}_{I_1, \dots, I_T} \left[ \|w_T^{(1,1)} - w_T^{(1,2)} - w_T^{(2,1)} + w_T^{(2,2)}\| \right] \\ & + O\left(\max_{j_1, k_1, j_2, k_2 \in \{1, 2\}} \mathbb{E}_{I_1, \dots, I_T} \left[ \|w_T^{(j_1, k_1)} - w_T^{(j_2, k_2)}\|^2 \right]\right). \end{aligned}$$

This completes the proof.  $\square$

**Lemma 12.** Under Assumption 3, if the learning rate satisfies  $\alpha_t \leq \frac{c}{t}$ , then for iterates  $w_t^{(a)}$  and  $w_t^{(b)}$  (trained on data sequences  $z^{(a)}$  and  $z^{(b)}$  of effective size  $N = s + 1$  that differ by one point):

$$\mathbb{E}_{I_1, \dots, I_t} \|w_t^{(a)} - w_t^{(b)}\| \leq \frac{2L}{\beta N} t^{c\beta}.$$

*Proof of Lemma 12.* Let  $w_t^{(a)}$  and  $w_t^{(b)}$  be the iterates at step  $t$  when training on data sequences  $z^{(a)}$  and  $z^{(b)}$  respectively. These sequences are of length  $N = s + 1$  and differ only at the  $N$ -th position (i.e.,  $z_N^{(a)} = z_a$  and  $z_N^{(b)} = z_b$ , while  $z_j^{(a)} = z_j^{(b)}$  for  $j < N$ ). The SGD update implies  $w_t^{(a)} = w_{t-1}^{(a)} - \frac{\alpha_t}{m} \sum_{j \in I_t} \nabla_w \ell(w_{t-1}^{(a)}; z_j^{(a)})$ .

The difference in iterates is:

$$\begin{aligned} \|w_t^{(a)} - w_t^{(b)}\| &= \left\| \left[ w_{t-1}^{(a)} - \frac{\alpha_t}{m} \sum_{j \in I_t} \nabla_w \ell(w_{t-1}^{(a)}; z_j^{(a)}) \right] - \left[ w_{t-1}^{(b)} - \frac{\alpha_t}{m} \sum_{j \in I_t} \nabla_w \ell(w_{t-1}^{(b)}; z_j^{(b)}) \right] \right\| \\ &= \left\| (w_{t-1}^{(a)} - w_{t-1}^{(b)}) - \frac{\alpha_t}{m} \sum_{j \in I_t} \left( \nabla_w \ell(w_{t-1}^{(a)}; z_j^{(a)}) - \nabla_w \ell(w_{t-1}^{(b)}; z_j^{(b)}) \right) \right\| \\ &\leq \|w_{t-1}^{(a)} - w_{t-1}^{(b)}\| + \frac{\alpha_t}{m} \sum_{j \in I_t} \left\| \nabla_w \ell(w_{t-1}^{(a)}; z_j^{(a)}) - \nabla_w \ell(w_{t-1}^{(b)}; z_j^{(b)}) \right\|. \quad (\text{by triangle inequality}) \end{aligned}$$

Let  $E_{t-1} = \|w_{t-1}^{(a)} - w_{t-1}^{(b)}\|$ . Consider the term  $\left\| \nabla_w \ell(w_{t-1}^{(a)}; z_j^{(a)}) - \nabla_w \ell(w_{t-1}^{(b)}; z_j^{(b)}) \right\|$  for  $j \in I_t$ .

Case 1: The data point  $z_j$  is common to both sequences. If  $z_j^{(a)} = z_j^{(b)} = z_j$  (i.e.,  $j < N$  or the point at index  $j$  in the sequences is from the common part  $\mathcal{S}$ ), then by  $\beta$ -smoothness of  $\ell$  (Assumption 3):

$$\left\| \nabla_w \ell(w_{t-1}^{(a)}; z_j) - \nabla_w \ell(w_{t-1}^{(b)}; z_j) \right\| \leq \beta \|w_{t-1}^{(a)} - w_{t-1}^{(b)}\| = \beta E_{t-1}.$$

Case 2: The data point  $z_j$  is the one that differs. If  $z_j^{(a)} = z_a$  and  $z_j^{(b)} = z_b$  (i.e.,  $j = N$ , the differing  $(s+1)$ -th point), then by Lipschitz assumption in Assumption 3,  $\|\nabla_w \ell(w, z)\|$  is bounded by  $L$ , so that,

$$\begin{aligned} \left\| \nabla_w \ell(w_{t-1}^{(a)}; z_a) - \nabla_w \ell(w_{t-1}^{(b)}; z_b) \right\| &\leq \left\| \nabla_w \ell(w_{t-1}^{(a)}; z_a) - \nabla_w \ell(w_{t-1}^{(b)}; z_a) \right\| \\ &\quad + \left\| \nabla_w \ell(w_{t-1}^{(b)}; z_a) - \nabla_w \ell(w_{t-1}^{(b)}; z_b) \right\| \\ &\leq \beta \left\| w_{t-1}^{(a)} - w_{t-1}^{(b)} \right\| + (\|\nabla_w \ell(w_{t-1}^{(b)}; z_a)\| + \|\nabla_w \ell(w_{t-1}^{(b)}; z_b)\|) \\ &\leq \beta E_{t-1} + 2L. \end{aligned}$$

Then the sum  $\sum_{j \in I_t} \left\| \nabla_w \ell(w_{t-1}^{(a)}; z_j^{(a)}) - \nabla_w \ell(w_{t-1}^{(b)}; z_j^{(b)}) \right\|$  can be bounded: It consists of  $(m - \mathbb{I}(N \in I_t))$  terms of Case 1 and  $\mathbb{I}(N \in I_t)$  terms of Case 2. Sum  $\leq (m - \mathbb{I}(N \in I_t))\beta E_{t-1} + \mathbb{I}(N \in I_t)(\beta E_{t-1} + 2L) = m\beta E_{t-1} + 2L\mathbb{I}(N \in I_t)$ . So,

$$\begin{aligned} \left\| w_t^{(a)} - w_t^{(b)} \right\| &\leq E_{t-1} + \frac{\alpha_t}{m} (m\beta E_{t-1} + 2L \cdot \mathbb{I}(N \in I_t)) \\ &= (1 + \alpha_t \beta) E_{t-1} + \frac{2\alpha_t L}{m} \cdot \mathbb{I}(N \in I_t) \\ &= (1 + \alpha_t \beta) \left\| w_{t-1}^{(a)} - w_{t-1}^{(b)} \right\| + \frac{2\alpha_t L}{m} \cdot \mathbb{I}(N \in I_t). \end{aligned}$$

This recurrence relation for the norm (holding for each realization of  $I_t$ ) leads to:

$$\left\| w_t^{(a)} - w_t^{(b)} \right\| \leq \sum_{t_0=1}^t \frac{2\alpha_{t_0} L}{m} \mathbb{I}(N \in I_{t_0}) \prod_{t'=t_0+1}^t (1 + \alpha_{t'} \beta). \quad (18)$$

Taking expectation over the mini-batch choices  $I_1, \dots, I_t$ , and substituting  $\alpha_t \leq \frac{c}{t}$ :

$$\begin{aligned} \mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\| &\leq \sum_{t_0=1}^t \frac{2\alpha_{t_0} L}{m} \mathbb{E}[\mathbb{I}(N \in I_{t_0})] \prod_{t'=t_0+1}^t (1 + \alpha_{t'} \beta) \\ &\leq \sum_{t_0=1}^t \frac{2cL}{t_0 m} \cdot \mathbb{P}(N \in I_{t_0}) \prod_{t'=t_0+1}^t \left( 1 + \frac{c\beta}{t'} \right). \end{aligned}$$

Since the mini-batch  $I_{t_0}$  of size  $m$  is sampled uniformly from the  $N = s+1$  data points,  $\mathbb{P}(N \in I_{t_0}) = \frac{m}{N} = \frac{m}{s+1}$ .

$$\begin{aligned} \mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\| &\leq \sum_{t_0=1}^t \frac{2cL}{t_0 m} \cdot \frac{m}{s+1} \cdot \exp \left( c\beta \sum_{t'=t_0+1}^t \frac{1}{t'} \right) \quad (\text{using } 1+x \leq e^x) \\ &\leq \frac{2cL}{s+1} \sum_{t_0=1}^t \frac{1}{t_0} \exp \left( c\beta \log \left( \frac{t}{t_0} \right) \right) \quad \left( \text{using } \sum_{k=a+1}^b 1/k \approx \log(b/a) \right) \\ &= \frac{2cL}{s+1} \sum_{t_0=1}^t \frac{1}{t_0} \left( \frac{t}{t_0} \right)^{c\beta}. \end{aligned}$$

Using the bound  $\sum_{k=1}^n \frac{1}{k} \left( \frac{n}{k} \right)^\gamma \leq \frac{n^\gamma}{\gamma}$  for  $\gamma > 0$  (here  $\gamma = c\beta$ ,  $n = t$ ,  $k = t_0$ ):

$$\mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\| \leq \frac{2cL}{s+1} \cdot \frac{t^{c\beta}}{c\beta} = \frac{2L}{(s+1)\beta} t^{c\beta}.$$

This completes the proof.  $\square$

**Lemma 13.** Under Assumption 3, if the learning rate satisfies  $\alpha_t \leq \frac{c}{t}$ , then for iterates  $w_t^{(a)}$  and  $w_t^{(b)}$  (trained on data sequences  $z^{(a)}$  and  $z^{(b)}$  of effective size  $N = s+1$  that differ by one point):

$$\mathbb{E}_{I_1, \dots, I_t} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2 \leq \frac{4L^2}{N} t^{2c\beta} \left( \frac{1}{N\beta^2} + \frac{2c^2}{m} \right).$$

*Proof of Lemma 13.* Let  $N = s + 1$  be the effective size of the data sequences  $z^{(a)}$  and  $z^{(b)}$ . From Equation (18) in the proof of Lemma 12, we have the bound for each realization of mini-batch choices:

$$\left\| w_t^{(a)} - w_t^{(b)} \right\| \leq X_t := \sum_{t_0=1}^t \frac{2\alpha_{t_0}L}{m} \mathbb{I}(N \in I_{t_0}) \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta).$$

We want to bound  $\mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2$ . Since  $\left\| w_t^{(a)} - w_t^{(b)} \right\| \geq 0$ , if  $\left\| w_t^{(a)} - w_t^{(b)} \right\| \leq X_t$ , then  $\left\| w_t^{(a)} - w_t^{(b)} \right\|^2 \leq X_t^2$ . Thus,  $\mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2 \leq \mathbb{E}[X_t^2]$ . Using the property  $\mathbb{E}[X_t^2] = (\mathbb{E}[X_t])^2 + \text{Var}[X_t]$ :

$$\begin{aligned} \mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2 &\leq (\mathbb{E}[X_t])^2 + \text{Var}[X_t] \\ &= \left( \mathbb{E} \left[ \sum_{t_0=1}^t \frac{2\alpha_{t_0}L}{m} \mathbb{I}(N \in I_{t_0}) \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta) \right] \right)^2 \\ &\quad + \text{Var} \left[ \sum_{t_0=1}^t \frac{2\alpha_{t_0}L}{m} \mathbb{I}(N \in I_{t_0}) \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta) \right]. \end{aligned}$$

Let  $Y_{t_0} = \frac{2\alpha_{t_0}L}{m} \mathbb{I}(N \in I_{t_0}) \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta)$ . Since the random variables  $\mathbb{I}(N \in I_{t_0})$  are independent across different time steps  $t_0$  (as mini-batches  $I_{t_0}$  are sampled independently), the terms  $Y_{t_0}$  are independent. Thus, the variance of the sum is the sum of variances:

$$\text{Var} \left[ \sum_{t_0=1}^t Y_{t_0} \right] = \sum_{t_0=1}^t \text{Var}(Y_{t_0}).$$

Let  $K_{t_0} = \frac{2\alpha_{t_0}L}{m} \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta)$ . This term is deterministic once  $\alpha$  values are fixed. Let  $p_{t_0} = \mathbb{P}(N \in I_{t_0}) = \frac{m}{N} = \frac{m}{s+1}$ . Then  $\text{Var}(Y_{t_0}) = \text{Var}(K_{t_0} \mathbb{I}(N \in I_{t_0})) = K_{t_0}^2 \text{Var}(\mathbb{I}(N \in I_{t_0})) = K_{t_0}^2 p_{t_0} (1 - p_{t_0}) \leq K_{t_0}^2 p_{t_0}$ . Using this observation for the variance term and the bound from Lemma 12 for the  $(\mathbb{E}[X_t])^2$  term:

$$\begin{aligned} \mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2 &\leq \left( \frac{2L}{(s+1)\beta} t^{c\beta} \right)^2 + \sum_{t_0=1}^t \left( \frac{2\alpha_{t_0}L}{m} \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta) \right)^2 \frac{m}{s+1} \\ &= \frac{4L^2}{(s+1)^2\beta^2} t^{2c\beta} + \sum_{t_0=1}^t \frac{4\alpha_{t_0}^2 L^2}{m^2} \cdot \frac{m}{s+1} \left( \prod_{t'=t_0+1}^t (1 + \alpha_{t'}\beta) \right)^2 \\ &\leq \frac{4L^2}{(s+1)^2\beta^2} t^{2c\beta} + \sum_{t_0=1}^t \frac{4c^2 L^2}{t_0^2 m (s+1)} \left( \prod_{t'=t_0+1}^t \exp\left(\frac{c\beta}{t'}\right) \right)^2 \\ &\leq \frac{4L^2}{(s+1)^2\beta^2} t^{2c\beta} + \sum_{t_0=1}^t \frac{4c^2 L^2}{t_0^2 m (s+1)} \exp\left(2c\beta \sum_{t'=t_0+1}^t \frac{1}{t'}\right) \\ &\leq \frac{4L^2}{(s+1)^2\beta^2} t^{2c\beta} + \frac{4c^2 L^2}{m(s+1)} \sum_{t_0=1}^t \frac{1}{t_0^2} \left( \frac{t}{t_0} \right)^{2c\beta}. \end{aligned}$$

Using the bound  $\sum_{k=1}^{\infty} \frac{1}{k^{2+\gamma}} \leq (1 + \frac{1}{1+\gamma})$  for  $\gamma \geq 0$  (here,  $k = t_0$ ,  $\gamma = 2c\beta$ ):

$$\begin{aligned} \mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2 &\leq \frac{4L^2}{(s+1)^2\beta^2} t^{2c\beta} + \frac{4c^2 L^2}{m(s+1)} \left( 1 + \frac{1}{1+2c\beta} \right) t^{2c\beta} \\ &\leq \frac{4L^2}{s+1} t^{2c\beta} \left( \frac{1}{(s+1)\beta^2} + \frac{2c^2}{m} \right) \end{aligned}$$

Replacing  $s + 1$  with  $N$  (the effective dataset size as per the lemma statement):

$$\mathbb{E} \left\| w_t^{(a)} - w_t^{(b)} \right\|^2 \leq \frac{4L^2}{N} t^{2c\beta} \left( \frac{1}{N\beta^2} + \frac{2c^2}{m} \right).$$

This completes the proof.  $\square$

**Lemma 14.** *Under Assumption 3, the loss function  $\ell(w; z)$  satisfies the following second-order approximation bound for all  $w, h \in \mathbb{R}^d$  and  $z \in \mathcal{Z}$ :*

$$\left\| \nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) - \nabla_w^2 \ell(w; z) h \right\| \leq \rho \|h\|^2.$$

*Proof of Lemma 14.* Define the function  $\phi(t) := \nabla_w \ell(w + th; z)$  for  $t \in [0, 1]$ . Then, the function  $\phi : [0, 1] \rightarrow \mathbb{R}^d$  is continuously differentiable. By the mean-value-type expansion result of McLeod [11], we can write:

$$\phi(1) = \phi(0) + \sum_{k=1}^d \lambda_k \phi'(t_k),$$

for some  $t_k \in (0, 1)$ , non-negative weights  $\lambda_k \geq 0$  such that  $\sum_{k=1}^d \lambda_k = 1$ .

The derivative  $\phi'(t)$  is  $\frac{d}{dt} \nabla_w \ell(w + th; z) = \nabla_w^2 \ell(w + th; z) h$ . Substituting the definitions of  $\phi(0)$ ,  $\phi(1)$ , and  $\phi'(t_k)$  into the expansion, we get:

$$\begin{aligned} \nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) &= \sum_{k=1}^d \lambda_k (\nabla_w^2 \ell(w + t_k h; z) h) \\ &= \left( \sum_{k=1}^d \lambda_k \nabla_w^2 \ell(w + t_k h; z) \right) h. \end{aligned}$$

To isolate the term  $\nabla_w^2 \ell(w; z) h$ , we add and subtract it:

$$\nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) = \nabla_w^2 \ell(w; z) h + \left( \sum_{k=1}^d \lambda_k [\nabla_w^2 \ell(w + t_k h; z) - \nabla_w^2 \ell(w; z)] \right) h.$$

Rearranging the terms, we have:

$$\nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) - \nabla_w^2 \ell(w; z) h = \sum_{k=1}^d \lambda_k [\nabla_w^2 \ell(w + t_k h; z) - \nabla_w^2 \ell(w; z)] h.$$

Taking norms on both sides and applying the triangle inequality for sums, followed by the properties of matrix norms:

$$\begin{aligned} \left\| \nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) - \nabla_w^2 \ell(w; z) h \right\| &\leq \left\| \sum_{k=1}^d \lambda_k [\nabla_w^2 \ell(w + t_k h; z) - \nabla_w^2 \ell(w; z)] h \right\| \\ &\leq \sum_{k=1}^d \lambda_k \left\| \nabla_w^2 \ell(w + t_k h; z) - \nabla_w^2 \ell(w; z) \right\| \cdot \|h\|. \end{aligned}$$

By the assumption of Lipschitz continuity of the Hessian of  $\ell$  with constant  $\rho$  in Assumption 3:

$$\left\| \nabla_w^2 \ell(w + t_k h; z) - \nabla_w^2 \ell(w; z) \right\| \leq \rho \|(w + t_k h) - w\| = \rho t_k \|h\|.$$

Substituting this into the inequality:

$$\begin{aligned} \left\| \nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) - \nabla_w^2 \ell(w; z) h \right\| &\leq \sum_{k=1}^d \lambda_k \cdot (\rho t_k \|h\|) \cdot \|h\| \\ &= \rho \|h\|^2 \sum_{k=1}^d \lambda_k t_k. \end{aligned}$$

Since  $t_k \in (0, 1)$ , we have  $t_k \leq 1$ . Also,  $\lambda_k \geq 0$  and  $\sum_{k=1}^d \lambda_k = 1$ . Therefore,  $\sum_{k=1}^d \lambda_k t_k \leq \sum_{k=1}^d \lambda_k \cdot 1 = 1$ . Thus,

$$\left\| \nabla_w \ell(w + h; z) - \nabla_w \ell(w; z) - \nabla_w^2 \ell(w; z) h \right\| \leq \rho \|h\|^2.$$

This completes the proof.  $\square$

### A.7 Proof of Proposition 3 (algorithmic stability of Influence Function (IF))

The proof of Proposition 3 uses some familiar techniques from the standard IF theory, but also features our original analysis. We first define an auxiliary parameter vector  $\tilde{\theta}_S$ . This vector is the minimizer of the empirical loss over  $S$  but with its denominator scaled as if two additional points were present in the averaging, which provides a good anchor point that facilitates a cleaner Taylor expansion form for  $\hat{\theta}_{S \cup \{z_1, z_2\}}$ .

$$\tilde{\theta}_S := \arg \min_{\theta} \left\{ \frac{1}{s+2} \sum_{z \in S} \ell(\theta; z) + \frac{\lambda}{2} \|\theta\|_2^2 \right\}.$$

Next, we present two technical lemmas, based on which, the validity of Proposition 3 would become clear. The proofs of these technical lemmas are relegated Appendix A.7.1.

**Lemma 15** (Expansion of parameter difference via IF). *Suppose the loss function  $\ell(\theta; z)$  is convex in  $\theta$ , three times continuously differentiable with respect to  $\theta$ , and its first-, second-, and third-order derivatives with respect to  $\theta$  are uniformly bounded for all  $z \in \mathcal{Z}$ . For any dataset  $S \in \mathcal{Z}^s$  (where  $s = |S|$ ) and any two data points  $z_1, z_2 \in \mathcal{Z}$ , let  $\tilde{\theta}_{S \cup \{z_1, z_2\}}$  be the parameter vector minimizing the regularized loss on  $S \cup \{z_1, z_2\}$  as per (13). Then we have*

1.  $\|\hat{\theta}_{S \cup \{z_1, z_2\}} - \tilde{\theta}_S\| = O(s^{-1})$ .
2.  $\hat{\theta}_{S \cup \{z_1, z_2\}} - \tilde{\theta}_S = -H_{\tilde{\theta}_S}^{-1} \left[ \nabla_{\theta} \ell(\tilde{\theta}_S; z_1) + \nabla_{\theta} \ell(\tilde{\theta}_S; z_2) \right] \frac{1}{s+2} + O(s^{-2})$ , where  $H_{\tilde{\theta}_S} := \left( \frac{1}{s+2} \sum_{z \in S} \nabla_{\theta}^2 \ell(\tilde{\theta}_S; z) \right) + \lambda I$ .

**Lemma 16** (Expansion of utility). *Suppose the performance metric  $u(\theta)$  is once continuously differentiable and its gradient  $\nabla_{\theta} u(\theta)$  is  $L_u$ -Lipschitz continuous and bounded. Then, for any parameter vectors  $\theta_A$  and  $\theta_B$ , we have*

$$u(\theta_A) - u(\theta_B) = \langle \nabla_{\theta} u(\theta_B), \theta_A - \theta_B \rangle + O(\|\theta_A - \theta_B\|^2).$$

*Proof of Proposition 3.* Let  $\Delta_U = U(S \cup \{z_1, z'_1\}) - U(S \cup \{z_1, z_2\}) - U(S \cup \{z'_1, z'_2\}) + U(S \cup \{z_2, z'_2\})$ . Recall  $U(\mathcal{X}) = u(\hat{\theta}_{\mathcal{X}})$ . We use Lemma 16 to expand each  $U(S \cup \{a, b\})$  around  $u(\tilde{\theta}_S)$ :

$$U(S \cup \{a, b\}) - u(\tilde{\theta}_S) = \langle \nabla_{\theta} u(\tilde{\theta}_S), \hat{\theta}_{S \cup \{a, b\}} - \tilde{\theta}_S \rangle + O(\|\hat{\theta}_{S \cup \{a, b\}} - \tilde{\theta}_S\|^2).$$

Let  $\delta\hat{\theta}_{ab} := \hat{\theta}_{S \cup \{a, b\}} - \tilde{\theta}_S$ . Lemma 15 implies that  $\|\delta\hat{\theta}_{ab}\|^2 = O(s^{-2})$ . It also states that the first-order component of  $\delta\hat{\theta}_{ab}$  is:

$$\delta\hat{\theta}_{ab}^{(1)} := -H_{\tilde{\theta}_S}^{-1} \left[ \nabla_{\theta} \ell(\tilde{\theta}_S; a) + \nabla_{\theta} \ell(\tilde{\theta}_S; b) \right] \frac{1}{s+2}.$$

Thus,  $\delta\hat{\theta}_{ab} = \delta\hat{\theta}_{ab}^{(1)} + O(s^{-2})$ . Substituting this into the expansion of  $U(S \cup \{a, b\}) - u(\tilde{\theta}_S)$  yields

$$U(S \cup \{a, b\}) = u(\tilde{\theta}_S) + \langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{ab}^{(1)} \rangle + \langle \nabla_{\theta} u(\tilde{\theta}_S), O(s^{-2}) \rangle + O(s^{-2}).$$

Since  $\nabla_{\theta} u(\tilde{\theta}_S)$  is bounded, the term  $\langle \nabla_{\theta} u(\tilde{\theta}_S), O(s^{-2}) \rangle$  is also  $O(s^{-2})$ . Therefore,  $U(S \cup \{a, b\}) = u(\tilde{\theta}_S) + \langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{ab}^{(1)} \rangle + O(s^{-2})$ .

Now, we substitute this expansion into the expression of  $\Delta_U$  and obtain

$$\begin{aligned} \Delta_U &= \left( u(\tilde{\theta}_S) + \langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{z_1 z'_1}^{(1)} \rangle \right) - \left( u(\tilde{\theta}_S) + \langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{z_1 z_2}^{(1)} \rangle \right) \\ &\quad - \left( u(\tilde{\theta}_S) + \langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{z'_1 z'_2}^{(1)} \rangle \right) + \left( u(\tilde{\theta}_S) + \langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{z_2 z'_2}^{(1)} \rangle \right) + \sum O(s^{-2}). \end{aligned}$$

The  $u(\tilde{\theta}_S)$  terms cancel. The sum of the four  $O(s^{-2})$  remainder terms is still  $O(s^{-2})$ . The sum of the first-order inner product terms is:

$$\langle \nabla_{\theta} u(\tilde{\theta}_S), \delta\hat{\theta}_{z_1 z'_1}^{(1)} - \delta\hat{\theta}_{z_1 z_2}^{(1)} - \delta\hat{\theta}_{z'_1 z'_2}^{(1)} + \delta\hat{\theta}_{z_2 z'_2}^{(1)} \rangle.$$

Let  $h_x = -\frac{1}{s+2}H_{\tilde{\theta}_S}^{-1}\nabla_{\theta}\ell(\tilde{\theta}_S; x)$ . Then  $\delta\tilde{\theta}_{ab}^{(1)} = h_a + h_b$ . The sum of influence terms becomes:

$$\begin{aligned} & \langle \nabla_{\theta}u(\tilde{\theta}_S), (h_{z_1} + h_{z'_1}) - (h_{z_1} + h_{z_2}) - (h_{z'_1} + h_{z'_2}) + (h_{z_2} + h_{z'_2}) \rangle \\ &= \langle \nabla_{\theta}u(\tilde{\theta}_S), h_{z_1} + h_{z'_1} - h_{z_1} - h_{z_2} - h_{z'_1} - h_{z'_2} + h_{z_2} + h_{z'_2} \rangle \\ &= \langle \nabla_{\theta}u(\tilde{\theta}_S), \mathbf{0} \rangle = 0. \end{aligned}$$

Thus, the first-order influence terms cancel out completely. The remaining terms are all  $O(s^{-2})$ . The proof of Proposition 3 is complete.  $\square$

### A.7.1 Lemmas used in the proof of Proposition 3

**Lemma 15** (Expansion of parameter difference via IF). *Suppose the loss function  $\ell(\theta; z)$  is convex in  $\theta$ , three times continuously differentiable with respect to  $\theta$ , and its first-, second-, and third-order derivatives with respect to  $\theta$  are uniformly bounded for all  $z \in \mathcal{Z}$ . For any dataset  $\mathcal{S} \in \mathcal{Z}^s$  (where  $s = |\mathcal{S}|$ ) and any two data points  $z_1, z_2 \in \mathcal{Z}$ , let  $\tilde{\theta}_{\mathcal{S} \cup \{z_1, z_2\}}$  be the parameter vector minimizing the regularized loss on  $\mathcal{S} \cup \{z_1, z_2\}$  as per (13). Then we have*

1.  $\|\tilde{\theta}_{\mathcal{S} \cup \{z_1, z_2\}} - \tilde{\theta}_S\| = O(s^{-1})$ .
2.  $\tilde{\theta}_{\mathcal{S} \cup \{z_1, z_2\}} - \tilde{\theta}_S = -H_{\tilde{\theta}_S}^{-1} \left[ \nabla_{\theta}\ell(\tilde{\theta}_S; z_1) + \nabla_{\theta}\ell(\tilde{\theta}_S; z_2) \right] \frac{1}{s+2} + O(s^{-2})$ , where  $H_{\tilde{\theta}_S} := \left( \frac{1}{s+2} \sum_{z \in \mathcal{S}} \nabla_{\theta}^2 \ell(\tilde{\theta}_S; z) \right) + \lambda I$ .

*Proof of Lemma 15.* For  $\delta_{\text{val}} \in [0, \frac{1}{s+2}]$ , define the perturbed objective function:

$$\mathcal{L}_{\text{pert}}(\theta, \delta_{\text{val}}) := \frac{1}{s+2} \sum_{z \in \mathcal{S}} \ell(\theta; z) + \delta_{\text{val}} [\ell(\theta; z_1) + \ell(\theta; z_2)] + \frac{\lambda}{2} \|\theta\|_2^2,$$

and its minimizer  $\hat{\theta}(\delta_{\text{val}}) := \arg \min_{\theta} \mathcal{L}_{\text{pert}}(\theta, \delta_{\text{val}})$ . By construction,  $\tilde{\theta}_S = \hat{\theta}(0)$ . The parameter vector  $\hat{\theta}_{\mathcal{S} \cup \{z_1, z_2\}}$  minimizes  $\frac{1}{s+2} \sum_{z \in \mathcal{S} \cup \{z_1, z_2\}} \ell(\theta; z) + \frac{\lambda}{2} \|\theta\|_2^2$ . This corresponds to  $\hat{\theta}(\frac{1}{s+2})$ . Thus,  $\hat{\theta}_{\mathcal{S} \cup \{z_1, z_2\}} - \tilde{\theta}_S = \hat{\theta}(\frac{1}{s+2}) - \hat{\theta}(0)$ .

The first-order condition for  $\hat{\theta}(\delta_{\text{val}})$  is  $F(\hat{\theta}(\delta_{\text{val}}), \delta_{\text{val}}) = 0$ , where

$$F(\theta, \delta_{\text{val}}) := \frac{\partial \mathcal{L}_{\text{pert}}}{\partial \theta} = \frac{1}{s+2} \sum_{z \in \mathcal{S}} \nabla_{\theta} \ell(\theta; z) + \delta_{\text{val}} [\nabla_{\theta} \ell(\theta; z_1) + \nabla_{\theta} \ell(\theta; z_2)] + \lambda \theta = 0.$$

Given the smoothness conditions on  $\ell$  (convexity, continuous third derivatives, bounded derivatives), the Higher-Order Implicit Function Theorem (e.g., Zorich [18]) ensures that  $\hat{\theta}(\delta_{\text{val}})$  is twice continuously differentiable with respect to  $\delta_{\text{val}}$  around  $\delta_{\text{val}} = 0$ . Differentiating  $F(\hat{\theta}(\delta_{\text{val}}), \delta_{\text{val}}) = 0$  w.r.t.  $\delta_{\text{val}}$  gives

$$\frac{\partial F}{\partial \theta} \frac{\partial \hat{\theta}}{\partial \delta_{\text{val}}} + \frac{\partial F}{\partial \delta_{\text{val}}} = 0.$$

Let

$$H(\delta_{\text{val}}) := \frac{\partial F}{\partial \theta} = \frac{1}{s+2} \sum_{z \in \mathcal{S}} \nabla_{\theta}^2 \ell(\hat{\theta}(\delta_{\text{val}}); z) + \delta_{\text{val}} [\nabla_{\theta}^2 \ell(\hat{\theta}(\delta_{\text{val}}); z_1) + \nabla_{\theta}^2 \ell(\hat{\theta}(\delta_{\text{val}}); z_2)] + \lambda I,$$

and

$$\frac{\partial F}{\partial \delta_{\text{val}}} = \nabla_{\theta} \ell(\hat{\theta}(\delta_{\text{val}}); z_1) + \nabla_{\theta} \ell(\hat{\theta}(\delta_{\text{val}}); z_2).$$

So,

$$\frac{\partial \hat{\theta}(\delta_{\text{val}})}{\partial \delta_{\text{val}}} = -H(\delta_{\text{val}})^{-1} \left[ \nabla_{\theta} \ell(\hat{\theta}(\delta_{\text{val}}); z_1) + \nabla_{\theta} \ell(\hat{\theta}(\delta_{\text{val}}); z_2) \right].$$

Due to convexity of  $\ell$ ,  $\nabla_{\theta}^2 \ell \succeq 0$ , so  $H(\delta_{\text{val}}) \succeq \lambda I$ . With  $\lambda > 0$ ,  $H(\delta_{\text{val}})$  is positive definite and its inverse is bounded. Uniform boundedness of  $\nabla_{\theta} \ell$  implies  $\frac{\partial \hat{\theta}(\delta_{\text{val}})}{\partial \delta_{\text{val}}}$  is uniformly bounded. Similarly, uniform boundedness of up to third-order derivatives of  $\ell$  ensures  $\frac{\partial^2 \hat{\theta}(\delta_{\text{val}})}{\partial \delta_{\text{val}}^2}$  is uniformly bounded.



By Taylor's theorem with Lagrange remainder, for some  $\xi \in [0, \frac{1}{s+2}]$ :

$$\hat{\theta}\left(\frac{1}{s+2}\right) - \hat{\theta}(0) = \frac{\partial \hat{\theta}(\delta_{\text{val}})}{\partial \delta_{\text{val}}}\bigg|_{\delta_{\text{val}}=0} \cdot \frac{1}{s+2} + \frac{1}{2} \frac{\partial^2 \hat{\theta}(\delta_{\text{val}})}{\partial \delta_{\text{val}}^2}\bigg|_{\delta_{\text{val}}=\xi} \cdot \left(\frac{1}{s+2}\right)^2.$$

Noting  $\hat{\theta}(0) = \tilde{\theta}_S$  and  $H(0) = H_{\tilde{\theta}_S}$  (as defined in the lemma statement), we have:

$$\hat{\theta}_{S \cup \{z_1, z_2\}} - \tilde{\theta}_S = -H_{\tilde{\theta}_S}^{-1} \left[ \nabla_{\theta} \ell(\tilde{\theta}_S; z_1) + \nabla_{\theta} \ell(\tilde{\theta}_S; z_2) \right] \frac{1}{s+2} + O\left(\frac{1}{(s+2)^2}\right).$$

This establishes part (2) of the lemma with the corrected negative sign. Part (1),  $\|\hat{\theta}_{S \cup \{z_1, z_2\}} - \tilde{\theta}_S\| = O(s^{-1})$ , follows because the first term is  $O(s^{-1})$  (since  $H_{\tilde{\theta}_S}^{-1}$  and gradients are bounded) and dominates the  $O(s^{-2})$  remainder for large  $s$ .  $\square$

**Lemma 16** (Expansion of utility). *Suppose the performance metric  $u(\theta)$  is once continuously differentiable and its gradient  $\nabla_{\theta} u(\theta)$  is  $L_u$ -Lipschitz continuous and bounded. Then, for any parameter vectors  $\theta_A$  and  $\theta_B$ , we have*

$$u(\theta_A) - u(\theta_B) = \langle \nabla_{\theta} u(\theta_B), \theta_A - \theta_B \rangle + O(\|\theta_A - \theta_B\|^2).$$

*Proof of Lemma 16.* By Taylor's theorem (or the Mean Value Theorem for vector functions), since  $u$  is once continuously differentiable, for some  $\xi$  on the line segment connecting  $\theta_A$  and  $\theta_B$ :

$$\begin{aligned} u(\theta_A) - u(\theta_B) &= \langle \nabla_{\theta} u(\xi), \theta_A - \theta_B \rangle \\ &= \langle \nabla_{\theta} u(\theta_B), \theta_A - \theta_B \rangle + \langle \nabla_{\theta} u(\xi) - \nabla_{\theta} u(\theta_B), \theta_A - \theta_B \rangle. \end{aligned}$$

The second term can be bounded using the Cauchy-Schwarz inequality and the  $L_u$ -Lipschitz continuity of  $\nabla_{\theta} u$ :

$$\begin{aligned} |\langle \nabla_{\theta} u(\xi) - \nabla_{\theta} u(\theta_B), \theta_A - \theta_B \rangle| &\leq \|\nabla_{\theta} u(\xi) - \nabla_{\theta} u(\theta_B)\|_2 \cdot \|\theta_A - \theta_B\|_2 \\ &\leq L_u \|\xi - \theta_B\|_2 \cdot \|\theta_A - \theta_B\|_2. \end{aligned}$$

Since  $\xi$  lies on the line segment between  $\theta_A$  and  $\theta_B$ , we have  $\|\xi - \theta_B\|_2 \leq \|\theta_A - \theta_B\|_2$ . Therefore, the absolute value of the second term is bounded by  $L_u \|\theta_A - \theta_B\|_2^2$ , which is  $O(\|\theta_A - \theta_B\|^2)$ . This establishes the result.  $\square$

## B Experimental Details

This appendix outlines the implementation details for all experiments in the main paper. To evaluate the effectiveness, efficiency, and robustness of FGSV, four experimental settings were employed. These consist of: (i) a synthetic data example in the introduction to illustrate the shell company attack (Section 1); (ii) a benchmark comparison on the SOU cooperative game (Section 4.1); (iii) an application to copyright attribution using a generative AI model with FlickrLogo-27 Dataset (Section 4.2); and (iv) an application to explainable AI using the Diabetes dataset (Section 4.3).

All experiments were conducted on the Unity high-performance computing cluster, provided by the College of Arts and Sciences at The Ohio State University. The copyright attribution experiment (Section 4.2) was executed on a GPU node equipped with an NVIDIA V100 GPU, 32GB VRAM, whereas all other experiments were conducted on CPU nodes featuring Intel Xeon E5-2699 v4 processors and 256 GB RAM. The code and instructions to reproduce the experiments are provided in the supplementary material and available at [https://github.com/KiljaeL/Faithful\\_GSV](https://github.com/KiljaeL/Faithful_GSV).

### B.1 Motivational Example (Section 1)

A synthetic setting is constructed to illustrate the shell company attack and its effect on group-level data valuation. We generate synthetic data for binary classification from a Gaussian mixture. Specifically, given a class label  $y_i \in \{0, 1\}$ , each sample  $x_i \in \mathbb{R}^2$  is drawn from  $\mathcal{N}(\mu_1, I_2)$  if  $y_i = 0$ , and from  $\mathcal{N}(\mu_2, I_2)$  if  $y_i = 1$ , where  $\mu_1 = [-3, 0]^\top$  and  $\mu_2 = [3, 0]^\top$ . The class label is sampled independently with equal probability  $p = 0.5$ . An additional  $n = 200$  samples are independently generated from the same distribution for testing.

The training data is partitioned into two equal-sized groups of 100 samples each. To simulate the shell company attack, the second group is further subdivided evenly into 2, 3, or 4 subgroups. This results in a total of  $k \in \{2, 3, 4, 5\}$  disjoint groups: the first group consistently contains 100 samples, while the remaining  $k - 1$  groups are allocated approximately  $100/(k - 1)$  samples each, with any remainder distributed to maintain balance (e.g., 33, 33, and 34 when  $k = 4$ ).

For each subset  $S$  of the training data, a logistic regression classifier is trained. The utility function  $U$  is defined as its classification accuracy on a held-out test set. If  $S$  contains only a single class, the utility is set to a constant value of  $1/2$ , corresponding to the expected accuracy of a random classifier.

While the experiment is qualitative and does not include confidence intervals, repeated runs with different random seeds produced consistent trends. FGSV is estimated using Algorithm 1 with threshold parameter  $\bar{s} = 20$  and Monte Carlo sample sizes  $m_1 = m_2 = 2000$ .

## B.2 Comparison with benchmark methods (Section 4.1)

For all methods compared here, we adjust their configurations, such that the total number of utility evaluations is aligned to the fixed number of 20,000. Note that since FGSV estimates the value for one group at a time, for fairness, the total budget is equally divided among the four groups, with 5,000 utility evaluations allocated per group. For FGSV, we set the threshold parameter to  $\bar{s} = 10$  and choose  $m_1 = m_2$  such that the total number of utility function calls sums to 20,000. Below, we describe the setup and computational structure of each method in detail, largely based on the implementation and descriptions from [9].

- **Permutation estimator [1, 4].**

The permutation estimator computes Shapley values by averaging marginal contribution across random permutations. At each iteration, it draws a permutation  $\pi$  uniformly at random from all possible orderings of the  $n$  elements and computes a Monte Carlo estimate based on the following alternative representation of the individual Shapley value:

$$SV(i) = \mathbb{E}_{\pi} [U(S_{\pi,i} \cup \{i\}) - U(S_{\pi,i})],$$

where  $S_{\pi,i}$  denotes the set of indices that appear before  $i$  in permutation  $\pi$ . The final estimate is obtained by averaging the marginal contributions over  $T$  sampled permutations:

$$\widehat{SV}_{\text{perm}}(i) = \frac{1}{T} \sum_{t=1}^T [U(S_{\pi_t,i} \cup i) - U(S_{\pi_t,i})],$$

where  $\pi_t$  denotes the  $t$ -th sample of random permutation. To simultaneously estimate  $SV(i)$  for all data points  $i \in [n]$  per permutation, this method requires  $n + 1$  utility evaluations – one for the empty set  $U(\emptyset)$ , and one each within incremental updates,  $U(\{\pi(1), \dots, \pi(j)\})$  for  $j = 1, \dots, n$ . Therefore, we set  $T = \lceil \frac{20000}{n+1} \rceil$  to maintain the total utility evaluation budget. For the last sampled permutation, the estimator stopped when the number of evaluations reached exactly 20,000.

- **Group Testing estimator [7, 15].**

This estimator computes pairwise differences of Shapley values rather than individual values. At each step, it draws a random size of subset  $s \in [n]$  with  $p(s) \propto \frac{1}{s} + \frac{1}{n-s+1}$ , and draws a random subset  $S_t \subseteq [n + 1]$  of size  $s$ , where  $(n + 1)$ -th element represents a dummy player, i.e., a player whose contribution is always zero and serves as a reference for baseline utility. In particular, for a given number of utility evaluations,  $T$ , a matrix  $B \in \mathbb{R}^{T \times (n+1)}$  is constructed, where the  $(t, i)$ -th entry is given by:

$$B_{t,i} = \begin{cases} U(S_t \setminus \{n + 1\}) & \text{if } i \in S_t, \\ 0 & \text{otherwise.} \end{cases}$$

We set  $T = 20000$ . Then, the Shapley value is approximated by

$$\widehat{SV}_{\text{gt}}(i) = \frac{Z}{T} \sum_{t=1}^T (B_{t,i} - B_{t,n+1}),$$

where  $B_{t,i}$  denotes the utility observed at iteration  $t$  when player  $i \in S_t$ , and  $Z = \sum_{s=1}^n p(s)$  is a normalization constant.

- **Complement Contribution estimator [17].**

The complement estimator approximates Shapley values by leveraging the symmetry between a subset and its complement. In particular, it relies on another equivalent representation of the Shapley value:

$$\text{SV}(i) = \frac{1}{n} \sum_{S \subseteq [n] \setminus \{i\}} \frac{U(S \cup \{i\}) - U([n] \setminus (S \cup \{i\}))}{\binom{n-1}{|S|}}.$$

This expression enables the estimator to reuse utility evaluations for both  $S$  and its complement  $[n] \setminus S$ . At each iteration, it samples a subset size  $s \in [n]$  uniformly and then uniformly samples a subset  $S \subseteq [n]$  of size  $s$ . Then, the final estimator is:

$$\widehat{\text{SV}}_{\text{cc}}(i) = \frac{1}{n} \sum_{s=1}^n \frac{1}{T_{i,s}} \sum_{t=1}^T [v_t \{\mathbb{I}(i \in S_t, |S_t| = s) - \mathbb{I}(i \notin S_t, |S_t| = n - s)\}],$$

where  $v_t = U(S_t) - U([n] \setminus S_t)$  and  $T_{i,s}$  is the number of such samples satisfying one of the two conditions. In our experiment, we set  $T = 20000/2 = 10000$  so that each iteration, which requires two utility evaluations, results in exactly 20,000 total evaluations.

- **One-for-All estimator [9].**

The One-for-All estimator is based on the following alternative formulation of the Shapley value<sup>3</sup>:

$$\text{SV}(i) = \frac{1}{n} \sum_{s=1}^n (\mathbb{E}_{S \subseteq [n], |S|=s, i \in S} [U(S)] - \mathbb{E}_{S \subseteq [n], |S|=s-1, i \notin S} [U(S)]),$$

This formulation allows each sampled subset to contribute to the estimates of all players, enabling efficient sample reuse. First, it deterministically allocates  $2n + 2$  utility evaluations for subset sizes  $s \in \{0, 1, n - 1, n\}$ , and uses them to compute the corresponding expectations exactly. Then, for the remaining possible subset sizes  $\{2, \dots, n - 2\}$ , it samples  $s$  with a predefined sampling probability  $q(s)$ , and then samples a subset  $S_t \subseteq [n]$  of size  $s$  uniformly at random. The final estimation is:

$$\begin{aligned} \widehat{\text{SV}}_{\text{ofa}}(i) &= \frac{1}{n} (U([n]) - U(\emptyset)) \\ &+ \frac{1}{n(n-1)} \left( \sum_{i \in S, |S|=n-1} U(S) - \sum_{i \notin S, |S|=1} U(S) \right) \\ &+ \frac{1}{n} \sum_{s=2}^{n-2} \left( \frac{1}{T_{i,s}^{\text{in}}} \sum_{t=1}^T U(S_t) \cdot \mathbb{I}(|S_t| = s, i \in S_t) - \frac{1}{T_{i,s}^{\text{out}}} \sum_{t=1}^T U(S_t) \cdot \mathbb{I}(|S_t| = s, i \notin S_t) \right), \end{aligned}$$

where  $T_{i,s}^{\text{in}}$  and  $T_{i,s}^{\text{out}}$  denote the number of Monte Carlo samples of size  $s$  in which  $i$  is included and excluded, respectively. In our experiment, we used  $q(s) \propto \frac{1}{\sqrt{s(n-s)}}$ , which is proved to be the optimal sampling distribution for Shapley value estimation. Next, after allocating  $2n + 2$  evaluations deterministically for subset sizes  $s \in \{0, 1, n - 1, n\}$ , the remaining evaluations are used for Monte Carlo sampling. That is, we set  $T = 20000 - (2n + 2)$  for sampling subset sizes  $s \in \{2, \dots, n - 2\}$ .

- **KernelSHAP [10].**

The Shapley value can be characterized as the solution to the following constrained optimization problem:

$$\text{SV} = \arg \min_{\phi \in \mathbb{R}^n} \sum_{\emptyset \subsetneq S \subsetneq [n]} w_S \left( U(S) - U(\emptyset) - \sum_{i \in S} \phi_i \right)^2, \quad \text{subject to} \quad \sum_{i=1}^n \phi_i = U([n]) - U(\emptyset),$$

<sup>3</sup>While the One-for-All estimator can estimate a variety of values (e.g., Banzhaf [14], Beta-Shapley [8]), we focus on its use for Shapley value estimation.

where  $w_S \propto \frac{1}{\binom{n-2}{|S|-1}}$  is the kernel weight assigned to each subset  $S$ . Based on this formulation, KernelSHAP constructs a sample-based approximation of the objective and solves the resulting weighted least squares problem to obtain an estimate of the Shapley value. In particular, at each iteration, it samples a subset size  $s \in \{1, \dots, n-1\}$  according to the distribution  $p(s) \propto \frac{1}{s(n-s)}$ , and then draws a subset  $S_t \subseteq [n]$  of size  $s$  uniformly at random. Each subset is encoded as a binary indicator vector  $\mathbf{1}_S \in \{0, 1\}^n$ , and the following quantities are updated:

$$\hat{A} = \frac{1}{T} \sum_{t=1}^T \mathbf{1}_{S_t} \mathbf{1}_{S_t}^\top, \quad \hat{b} = \frac{1}{T} \sum_{t=1}^T (U(S_t) - U(\emptyset)) \cdot \mathbf{1}_{S_t}.$$

The KernelSHAP estimator is obtained in closed form as:

$$\widehat{\text{SV}}_{\text{ks}} = \hat{A}^{-1} \left( \hat{b} - \frac{\mathbf{1}^\top \hat{A}^{-1} \hat{b} - U([n]) + U(\emptyset)}{\mathbf{1}^\top \hat{A}^{-1} \mathbf{1}} \cdot \mathbf{1} \right),$$

where  $\mathbf{1} \in \mathbb{R}^n$  denotes the vector with all-one entries. We set  $T = 20000$  in our implementation of KernelSHAP.

- **Unbiased KernelSHAP [2].**

The Unbiased KernelSHAP estimator is a variant of KernelSHAP that utilizes the fact that the exact gram matrix,  $A = \mathbb{E}[\mathbf{1}_S \mathbf{1}_S^\top]$ , admits a closed-form expression under the same distribution  $w_S \propto \frac{1}{\binom{n-2}{|S|-1}}$ . Instead of estimating  $A$  empirically from samples, this estimator directly computes  $A$ , where its  $(i, j)$ -entry is defined as:

$$A_{ij} = \begin{cases} \frac{1}{2} & \text{if } i = j, \\ \frac{1}{n(n-1)} \frac{\sum_{s=2}^{n-1} \frac{s-1}{n-s}}{\sum_{s=1}^{n-1} \frac{1}{s(n-s)}} & \text{otherwise.} \end{cases}$$

Unbiased KernelSHAP replaces the empirical matrix  $\hat{A}$  in KernelSHAP with its closed-form expectation  $A$ :

$$\widehat{\text{SV}}_{\text{uks}} = A^{-1} \left( \hat{b} - \frac{\mathbf{1}^\top A^{-1} \hat{b} - U([n]) + U(\emptyset)}{\mathbf{1}^\top A^{-1} \mathbf{1}} \cdot \mathbf{1} \right).$$

In our implementation, we set  $T = 20000$ .

- **LeverageSHAP [12].**

LeverageSHAP is another variant of KernelSHAP designed to reduce the variance of the estimator and improve computational efficiency. Unlike KernelSHAP, which samples subset sizes  $s \in \{1, \dots, n-1\}$  with the weighting  $p(s) \propto \frac{1}{s(n-s)}$ , LeverageSHAP draws  $s$  uniformly (i.e.,  $p(s) \propto 1$ ) and compensates for the mismatch via a correction factor  $w(s) = \sqrt{s(n-s)}$ . Also, it adopts paired sampling, which also includes complement  $\bar{S}_t = S_t^c$  in the sample pool when each subset  $S_t$  is sampled. Based on these modification, it computes the following quantities:

$$\begin{aligned} \tilde{A} &= \frac{1}{2T} \sum_{t=1}^T w(|S_t|) \left\{ \mathbf{1}_{S_t} \mathbf{1}_{S_t}^\top + \mathbf{1}_{\bar{S}_t} \mathbf{1}_{\bar{S}_t}^\top \right\} \\ \tilde{b} &= \frac{1}{2T} \sum_{t=1}^T w(|S_t|) \left\{ (U(S_t) - U(\emptyset)) \cdot \mathbf{1}_{S_t} + (U(\bar{S}_t) - U(\emptyset)) \cdot \mathbf{1}_{\bar{S}_t} \right\}. \end{aligned}$$

The final LeverageSHAP estimator is then computed by:

$$\widehat{\text{SV}}_{\text{ls}} = \tilde{A}^{-1} \left( \tilde{b} - \frac{\mathbf{1}^\top \tilde{A}^{-1} \tilde{b} - U([n]) + U(\emptyset)}{\mathbf{1}^\top \tilde{A}^{-1} \mathbf{1}} \cdot \mathbf{1} \right).$$

For LeverageSHAP, we set  $T = 20000/2 = 10000$  since each iteration requires two utility evaluations due to the use of paired sampling.

Although all methods use the same total number of utility evaluations (20,000), their actual runtimes differ substantially, as shown in the second row of Figure 2. These discrepancies are attributable to several algorithmic and implementation-specific factors, including computational complexity and sampling distribution over subset sizes.

First, **Group Testing** requires the construction of a utility matrix  $B \in \mathbb{R}^{T \times (n+1)}$ . Maintaining this large matrix introduces relatively high memory usage and per-iteration computational overhead, as each utility evaluation must be copied into multiple locations with structural indexing. In contrast, other estimators maintain only low-dimensional accumulators or regression statistics, which impose negligible memory cost.

Second, the cost of evaluating  $U(S)$  varies dramatically with the size of the subset,  $|S| = s$ . Consequently, the *subset size sampling distribution* required by each method significantly impacts speed. Among the benchmarks, **Group Testing**:  $p(s) \propto \frac{1}{s} + \frac{1}{n-s+1}$ ; **One-for-All**:  $q(s) \propto \frac{1}{\sqrt{s(n-s)}}$ ; and both **KernelSHAP** and **Unbiased KernelSHAP**:  $p(s) \propto \frac{1}{s(n-s)}$ . They all employ non-uniform subset size distributions that tend to oversample either very small or very large subsets. This matters significantly in the SOU game, where evaluating  $U(S)$  requires checking whether  $\mathcal{A}_j \subseteq S$  holds for all  $j \in [d]$ . Figure 5 empirically demonstrates the near-exponential growth in computation time for evaluating  $U(S)$  as a function of subset size  $s$ , across varying values of  $n \in \{64, 128, 256\}$ . This occurs because the time required for set inclusion checks (e.g., `issubset()` in Python) increases rapidly with subset size. As a result, estimators that disproportionately sample such extreme subset sizes tend to exhibit higher average computation time per evaluation compared to those that sample subset sizes uniformly (e.g., **Permutation**, **Complement Contribution**, **LeverageSHAP**). Thus, sampling behavior—not just the number of evaluations—directly impacts computational efficiency. On the other hand, **FGSV** does not sample subset sizes  $s$ , but instead performs estimation through an explicit loop over all possible values of  $s$ . As shown in Algorithm 2, when  $s < \bar{s}$ , it evaluates the utility function over all grid points  $s_1 \in [\max\{0, s + s_0 - n\}, \min\{s, s_0\}]$ . For  $s \geq \bar{s}$ , the approximation is performed using only two representative values of  $s_1$ , thereby significantly reducing the number of required evaluations. This thresholding mechanism ensures that, under a fixed budget of utility calls, FGSV concentrates more computation on smaller subset sizes, leading to lower overall runtime.

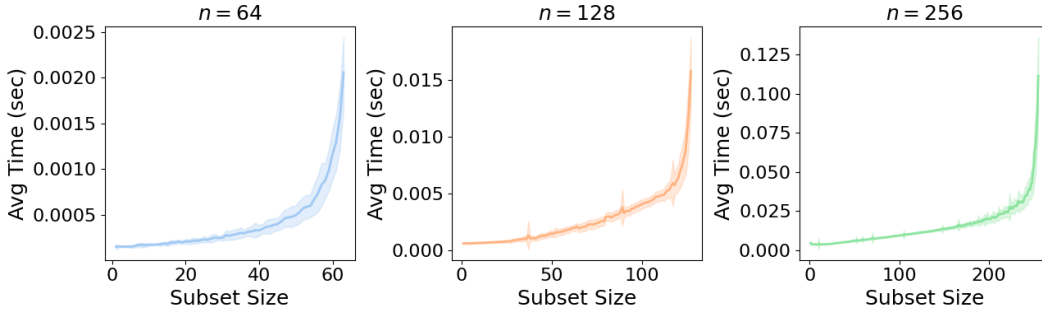


Figure 5: Empirical average runtime (in seconds) of  $U(S)$  evaluation as a function of subset size  $s = |S|$ . Each curve represents the mean over 50 randomly sampled subsets of size  $s$ ; shaded areas indicate  $\pm 1$  standard deviation.

Third, **KernelSHAP** and its variants—**Unbiased KernelSHAP** and **LeverageSHAP**—require solving a constrained least squares problem involving dense matrix inversions and multiplications. These linear algebra operations, performed either during or after sampling, dominate the overall computational cost, making these methods significantly slower than alternatives that rely solely on running averages or marginal contribution estimates.

### B.3 Faithful copyright attribution in generative AI (Section 4.2)

We evaluate SRS and FSRS in the context of group-level copyright attribution for generative models, using a logo generation task based on Stable Diffusion. To facilitate comparison, we mostly followed the experimental setup in [16]. Nonetheless, for clarity and completeness, we summarize the relevant details below.

To construct the experiment, we select four logo classes (Google, Sprite, Vodafone, and Starbucks) for evaluation, and use the remaining 23 brands to initialize a baseline model via fine-tuning of Stable Diffusion v1.4 [13]. We then assess the impact of fine-tuning with each of the four excluded brands by measuring how their inclusion alters the generation behavior of the model.

Fine-tuning is performed using Low-Rank Adaptation (LoRA) [6], a technique that inserts trainable low-rank matrices into the attention layers of the diffusion model. The rank  $r$  and scaling factor  $\alpha$  are both set to 8, and fine-tuning is only applied to the attention layers of the UNet module, while all other weights of the Stable Diffusion model are kept frozen. This configuration ensures efficient and stable adaptation to the small-scale dataset, in line with standard LoRA usage.<sup>4</sup>

We use a learning rate of 0.0001, a batch size of 4, and train for 10 epochs. After fine-tuning, we generate  $N_{\text{MC}} = 20$  images for each of the four selected brands using the prompt “A logo by [brand name].” Image generation is performed with 25 denoising steps and classifier-free guidance (scale 7.5), implemented using the DDPM Scheduler.

Following the formulation of [16], we define the utility function as the log-likelihood of generated images. For a given length of the denoising steps  $T$ , let  $(x_T, \dots, x_1, x_0)$  denote the reverse trajectory defined by the diffusion scheduler, where  $x_T \sim \mathcal{N}(0, I)$  is the initial latent noise, and  $x_0 = x_{(\text{gen})}$  is the final generated image. The likelihood of  $x_{(\text{gen})}$  under model parameters  $\theta$  is defined as:

$$p_{\theta}(x_{(\text{gen})}) = p_{\theta}(x_0) = \mathbb{E}_{x_1} [p_{\theta}(x_0 | x_1)],$$

by the Markov property of the diffusion process. The conditional distribution  $p_{\theta}(x_0 | x_1)$  is Gaussian, given by

$$p_{\theta}(x_0 | x_1) \sim \mathcal{N}\left(x_0; \frac{1}{\sqrt{\alpha_1}} \left(x_1^{(j)} - \sqrt{1 - \alpha_1} \hat{\epsilon}_{\theta}(x_1^{(j)}, 1)\right), \sigma_1^2 I\right),$$

where  $\hat{\epsilon}_{\theta}(x_t, t)$  denotes the predicted noise at step  $t$ , and  $\alpha_1, \sigma_1^2$  are scheduler-specific constants. Based on these formulation, the likelihood of  $x_{(\text{gen})}$  can be approximated as

$$\frac{1}{N_{\text{MC}}} \sum_{j=1}^{N_{\text{MC}}} \mathcal{N}\left(x_0; \frac{1}{\sqrt{\alpha_1}} \left(x_1^{(j)} - \sqrt{1 - \alpha_1} \hat{\epsilon}_{\theta}(x_1^{(j)}, 1)\right), \sigma_1^2 I\right),$$

where  $x_1^{(j)}$  for  $j = 1, \dots, N_{\text{MC}}$  is sampled by reversing the diffusion process from a standard Gaussian noise vector  $x_T^{(j)} \stackrel{iid}{\sim} \mathcal{N}(0, I)$ .

In our setting, the model is initialized from a pretrained checkpoint based on the remaining 23 brands, which stabilizes the fine-tuning process even when the input subset is small. This design parallels the data augmentation strategy described in Section 3.4, where non-informative examples are added to ensure that the utility function remains well-defined. Ultimately, the resulting setup can be conceptually viewed as one where the utility function consistently receives a sufficiently large dataset as input. Hence, the computation procedure can be understood as effectively a special case of Algorithm 2, with the difference that our computation injects *informative*, rather than non-informative, data points. As a result, when theoretically understanding the performance of our method here, we resort to Theorem 2, thinking that  $s$  is lower-bounded.

Finally, we used  $m = 2$  for this approximation. This choice was primarily informed by practical constraints: the computation for a single sample, including both fine-tuning and utility evaluation, takes nearly 30 hours given the limited hardware resources available to us. On the other hand, however, we observed that the FSRS estimates well-illustrates the *faithfulness* property of FGSV—the preservation of group-level value under further partitioning—even within each single experiment. As shown in Figure 3, which displays the result from a single experiment, the FSRS values remained stable before and after splitting the Google and Sprite brands. This suggests that the empirical evidence is clear even with a small  $m$ . To understand this phenomenon, we attribute such empirical stability to the fact that the diffusion model here was initialized by a pretrained baseline model. The baseline model was trained by on big data with very significant computing resources, which tends to yield rather stable initializations. Consequently, fine-tuning outcomes across random seeds tend to inherit some stability and would not vary wildly.

<sup>4</sup><https://huggingface.co/blog/lorax>

#### B.4 Faithful explainable AI (Section 4.3)

We evaluated FGSV and GSV on the Diabetes dataset [3], which includes 442 samples with 10 demographic and health-related features. The dataset is preprocessed by centering and standardizing the response variable and splitting the dataset into 350 training and 92 test samples using a fixed random seed.

For the base model, we use ridge regression with regularization strength  $\alpha = 0.01$ . The utility  $U(S)$  is defined as the negative mean squared error (MSE) on the test set; for empty subsets  $S$ , the utility is defined as the negative variance of the test labels, which corresponds to a baseline predictor that always outputs the test mean. FGSV is estimated using Algorithm 2 with parameters  $\bar{s} = 35$  and  $m_1 = m_2 = 1000$ , while GSV is computed exactly. Each setting is repeated for 30 times using independent Monte Carlo replications.

## C Approximation algorithm with non-informative data augmentation

We present the approximation algorithm with non-informative data augmentation in Algorithm 2.

---

**Algorithm 2** Approximate FGSV with non-informative data augmentation for small input sizes

---

**Require:** Dataset  $\mathcal{D}$ , group  $\mathcal{S}_0$ , size threshold  $B$ , subsample size  $m$ , non-informative distribution  $\mathcal{P}_{\text{null}}$ .

- 1: Initialize  $n \leftarrow |\mathcal{D}|$ ,  $s_0 \leftarrow |\mathcal{S}_0|$ ,  $\alpha_0 \leftarrow s_0/n$ .
- 2: Initialize total sum for correction terms  $\widehat{\mathcal{T}}_{\text{sum}} \leftarrow 0$ .
- 3: **for**  $s = 1$  to  $n - 1$  **do**
- 4:   Set  $s_1^* \leftarrow \lfloor s\alpha_0 \rfloor$ . {Expected intersection size for current  $s$ }
- 5:   Initialize sum of utility differences  $S_{\Delta U} \leftarrow 0$ .
- 6:   **for**  $j = 1$  to  $m$  **do**
- 7:     Sample a base tuple  $(\mathcal{S}, z_1, z_2)$  i.i.d. from  $\mathcal{B}_{s, s_1^*} = \{(\mathcal{S}, z_1, z_2) : \mathcal{S} \subseteq \mathcal{D}, |\mathcal{S}| = s, |\mathcal{S} \cap \mathcal{S}_0| = s_1, z_1 \in \mathcal{S}_0 \setminus \mathcal{S}, z_2 \in \mathcal{S}_0^c \setminus \mathcal{S}\}$ , where  $\mathcal{S}_0 = \{z_i : i \in \mathcal{S}_0\}$ .
- 8:     **if**  $s < B$  **then**
- 9:       Sample augmentation set  $\mathcal{S}_{\text{null}} = \{z'_1, \dots, z'_{B-s}\}$  where  $z'_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{P}_{\text{null}}$ .
- 10:        $U_1 \leftarrow U(\mathcal{S} \cup \{z_1\} \cup \mathcal{S}_{\text{null}})$ .
- 11:        $U_2 \leftarrow U(\mathcal{S} \cup \{z_2\} \cup \mathcal{S}_{\text{null}})$ .
- 12:     **else**
- 13:        $U_1 \leftarrow U(\mathcal{S} \cup \{z_1\})$ .
- 14:        $U_2 \leftarrow U(\mathcal{S} \cup \{z_2\})$ .
- 15:     **end if**
- 16:      $S_{\Delta U} \leftarrow S_{\Delta U} + (U_1 - U_2)$ .
- 17:   **end for**
- 18:    $\widehat{\Delta\mu} \left( \frac{s_1^*}{s}; s, s_0, n \right) \leftarrow \frac{1}{m} S_{\Delta U}$ .
- 19:    $\widehat{\mathcal{T}}(s) \leftarrow \frac{n}{n-1} \alpha_0 (1 - \alpha_0) \cdot \widehat{\Delta\mu} \left( \frac{s_1^*}{s}; s, s_0, n \right)$ .
- 20:    $\widehat{\mathcal{T}}_{\text{sum}} \leftarrow \widehat{\mathcal{T}}_{\text{sum}} + \widehat{\mathcal{T}}(s)$ .
- 21: **end for**
- 22:  $G_0 \leftarrow \frac{s_0}{n} [U([n]) - U(\emptyset)]$ .
- 23: **return**  $G_0 + \widehat{\mathcal{T}}_{\text{sum}}$ .

---

To theoretically justify Algorithm 2, we first define a padded utility function

$$\widetilde{U}(S) = \begin{cases} \mathbb{E}[U(S \cup S_{\text{null}})], & |S| < B, \\ U(S), & |S| \geq B, \end{cases}$$

where  $S_{\text{null}} \sim \mathcal{P}_{\text{null}}^{B-|S|}$ .

Under the assumption that  $U$  satisfies Assumptions 1 and 2 and is  $O(1/s)$ -deletion-stable for  $|S| \geq B$ , one can verify that  $\widetilde{U}$  retains these properties for all  $S$ . Applying Algorithm 2 to  $\widetilde{U}$  with the parameter choices

$$\begin{aligned} \bar{s} &\asymp \epsilon^{-1/v}, \quad m_1 \asymp \epsilon^{-\frac{2(2+v)}{v}} \log(n/\delta), \\ m_2 &\asymp \max \left\{ 1, \epsilon^{-2} (\alpha_0 (1 - \alpha_0))^2 \log^3(n/\delta) \right\}, \end{aligned}$$

yields an  $(\epsilon, \delta)$ -approximation of  $\widetilde{\text{FGSV}}(S_0)$  that requires

$$O \left( n \cdot \max \{ 1, \alpha_0 (1 - \alpha_0) \}^2 (\log n)^3 \right)$$

utility evaluations, matching the bounds in the original Theorem 3.



## References

- [1] J. Castro, D. Gómez, and J. Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.
- [2] I. Covert and S.-I. Lee. Improving kernelshap: Practical Shapley value estimation using linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3457–3465. PMLR, 2021.
- [3] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.
- [4] A. Ghorbani and J. Zou. Data Shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pages 2242–2251. PMLR, 09–15 Jun 2019.
- [5] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.
- [6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.
- [7] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. M. Gurel, B. Li, C. Zhang, C. Spanos, and D. Song. Efficient task specific data valuation for nearest neighbor algorithms. *Proceedings of the VLDB Endowment*, 12(11):1610–1623, 2018.
- [8] Y. Kwon and J. Zou. Beta Shapley: a unified and noise-reduced data valuation framework for machine learning. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151, pages 8780–8802, 2022.
- [9] W. Li and Y. Yu. One sample fits all: Approximating all probabilistic values simultaneously and efficiently. *arXiv preprint arXiv:2410.23808*, 2024.
- [10] S. Lundberg. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.
- [11] R. M. McLeod. Mean value theorems for vector valued functions. *Proceedings of the Edinburgh Mathematical Society*, 14(3):197–209, 1965.
- [12] C. Musco and R. T. Witter. Provably accurate Shapley value estimation via leverage score sampling. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [13] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.
- [14] J. T. Wang and R. Jia. Data Banzhaf: A robust data valuation framework for machine learning. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 6388–6421. PMLR, 25–27 Apr 2023.
- [15] J. T. Wang and R. Jia. A note on “Towards efficient data valuation based on the shapley value”. *arXiv preprint arXiv:2302.11431*, 2023.
- [16] J. T. Wang, Z. Deng, H. Chiba-Okabe, B. Barak, and W. J. Su. An economic solution to copyright challenges of generative ai. *arXiv preprint arXiv:2404.13964*, 2024.
- [17] J. Zhang, Q. Sun, J. Liu, L. Xiong, J. Pei, and K. Ren. Efficient sampling approaches to Shapley value approximation. *Proceedings of the ACM on Management of Data*, 1(1):1–24, 2023.
- [18] V. A. Zorich. *Mathematical Analysis II*, pages 41–108. Berlin, Heidelberg, 2016. ISBN 978-3-662-48993-2.