Dear Reviewer #1:

> The concept of submodular optimization with noisy (unbiased) oracles has been introduced before
Thanks for pointing out related works. As the reviewer mentioned, for submodular function *maximization* problems, there have been works considering noisy oracles. We will mention three papers you suggested in the revised version.

> Closing the gap between lower and upper bound.
Yes, this is a significant future work. This seems to be a difficult problem as it has been open to shave off polynomial gaps between lower and upper bounds in zeroth order convex optimization problems with noisy evaluation as well.

> Also, experiments showing that SGD with ... considerably faster in large-scale settings would be beneficial.
We agree with this suggestion. We are leaving empirical evaluation as an important future task.

Dear Reviewer #2:

Thanks for providing many comments and questions that help improve our paper. We will modify the manuscript by incorporating the following points:

> On line 178, the authors should acknowledge this construction of subgradient.
We will add the references you suggested in the revised version. The current manuscript mentions no specific literature because we consider that these expressions of subgradient are a sort of folklore and because the oldest literature is not clear. We can find a similar expression in Lemma 6.19 of the book by Fujishige.

> Using Lovasz extension and stochastic projected gradient descent are standard.
To our knowledge, the only literature that considers the combination of Lovasz extension and the *stochastic* projected gradient is [17], which is mentioned in lines 74-76, 86-96. Combining Lovasz extension and (exact) gradient methods can be found in more literature such as [4] and [5]. We will add a more detailed review in the revised version.

> On line 208, I was confused why I_t is a subset of 0, 1, ..., n.; This should be $\{1, 2, \ldots, n\}$. We will fix it.

> Furthermore, it is extremely confusing that the construction from line 207 to line 210 seems not to rely on x_t.
Because σ depends on x_t, $\hat{g}_t$ relies on x_t. We emphasize this in the revised version.

> On lines 217-218, the notation $\bigcup_{i \in J_t} \{S_\sigma(i), S_\sigma(i-1)\}$ is perplexing.; This means that the set of queries $\{X_t^{(j)}\}_{j=1}^k$ must include $S_\sigma(i)$ and $S_\sigma(i-1)$ for all $i \in J_t$, i.e., $\hat{f}_t(S_\sigma(i))$ and $\hat{f}_t(S_\sigma(i-1))$ must be evaluated for all $i \in J_t$.

> What is the high-level intuition that the estimator (16) is better?
As the reviewer pointed out, a key factor is that the vector $(f_t(S_\sigma(i)) - f_t(S_\sigma(i-1)))_{i=1}^n$ has a smaller norm than $(f_t(S_\sigma(i)))_{i=0}^n$, which is implied by Lemma 8 in [17] or Lemma 1 in [25]. We will modify the manuscript as suggested.

> the authors should justify in the paper that the assumption regarding multi-point evaluations is a realistic assumption.
For example, in the case of pricing optimization (lines 33-42) for E-commerce, we can get multiple-point feedback by employing the A/B-testing framework, i.e., by showing different prices to randomly divided groups of customers.

> What did the authors mean by "stochastically independent"? ; We mean usual independence.

> Why pick i_X and s_X? Are they used to make sure that evaluation of $\hat{f}$ is independent given different X?
Yes, we pick i_X and s_X for each X to let $\hat{f}(X)$ be independent for different X. This is needed in the proof of Lemma 4, to obtain a larger regret lower bound. (In line 254, $h_{i_X}(S^*)h_i(X)$ should be replaced with $h_{i_X}(S^*)h_{i_X}(X)$.)

> Why is this construction (sampled from $F'(S^*, \varepsilon)$) a submodular function? ... is it indeed a modular function?
The expectation $f_{S^*,\varepsilon}$ of $F'(S^*, \varepsilon)$ is indeed a modular function. (proof) From (19), $f_{S^*,\varepsilon}$ is a linear combination of $\{h_i\}_{i=1}^d$. Since an arbitrary linear combination of modular functions is modular as well, it suffices to show h_i is modular. From the definition of h_i (line 245), we can see that all $X, Y \subseteq [n]$ satisfy $h_i(X) + h_i(Y) = h_i(X \cup Y) + h_i(X \cap Y)$. In fact, both sides equal to -2 if $X, Y \ni i$ ($\Leftrightarrow X \cap Y \ni i$), to 2 if $X, Y \not\ni i$ ($\Leftrightarrow X \cup Y \not\ni i$), and to 0 otherwise.

> What is the connection between $\hat{f}$ on line 248 and $\hat{f}$ on line 254?
Both have the same expectation given in (19) but follow different distributions. Line 248 gives the definition of $\hat{f} \sim F(S^*, \varepsilon)$ and line 254 is for $\hat{f} \sim F'(S^*, \varepsilon)$. A difference of them is mentioned in lines 255-256. Besides, $f(X)$ are independent for different X if $f \sim F'(S^*, \varepsilon)$ while $F(S^*, \varepsilon)$ does not have this property.

Dear Reviewer #3:

> Line 172: The range of f is missing; Thanks for pointing out the typo. We will fix it in the revised version.

> It would be nice if there are experimental results that confirm the effectiveness of the proposed method.
We agree with your suggestion. We consider the empirical evaluation of our method to be an important future work.