A SUPPLEMENTARY LEMMAS

Lemma 7. *The clipped error ℓ_c satisfies the triangle inequality.*

Proof: This results follows because $|x|_c = |x{-}y{+}y|_c \le |x{-}y|_c{+}|y|_c$, still holds under clipping. To see why, consider the following. If either $|x-y|_c$ or $|y|_c$ are clipped to c, then clearly the sum is larger than $|x|_c$. Otherwise, if only $|x|_c$ is clipped to c, then it can only have been strictly decreased and again the inequality must hold. Once we have this inequality, we can use the fact that $|x| \le |x-y| + |y|$ and $|y| \le |x-y| + |x|$ to get the $|x| - |y| \le |x-y|$ and $|y| - |x| \le |x-y|$. ■

Lemma 6. For a state $s \in \mathcal{S}$, Algorithm 2 returns an ϵ, δ, τ-approximation $\bar{v}(s)$:

$$|\bar{v}^*(s) - \bar{v}(s)| \le \epsilon|\bar{v}^*(s)| + \epsilon\tau \tag{8}$$

Proof: We follow a similar argument to Mnih et al. [2008, Section 3.1]. The empirical Bernstein bound [Audibert et al., 2007] states that, for a sample average $\bar{g}_t = \frac{1}{t}\sum_{j=1}^{t} g_j$ of t unbiased samples g_j

$$|\bar{v}^*(s) - \bar{g}_t| \le c_t$$

where

$$c_t = \bar{\sigma}_t^{(r)}\sqrt{\frac{2\log(3/\delta)}{t}} + 3\log(3/\delta)\frac{V_{\max}}{t} \tag{18}$$

$$\bar{\sigma}_t^{(r)} = \sqrt{\frac{1}{t}\sum_{j=1}^{t}(g_j - \bar{v}(s))^2} \tag{19}$$

Algorithm 2 estimates lower and upper bounds, based on this concentration inequality, guaranteeing that the absolute value of the true value is between these bounds with probability at least $1-\delta$. Algorithm 2 terminates when either of the following cases are satisfied:

[Case : 1] $(1+\epsilon)\text{LB} + 2\epsilon\tau \ge (1-\epsilon)\text{UB}$ and returns $\bar{v} = \frac{\text{sign}(\bar{g}_t)}{2}((1+\epsilon)\text{LB} + (1-\epsilon)\text{UB})$.

[Case : 2] $\frac{\widehat{\text{UB}}-\widehat{\text{LB}}}{2} \le \epsilon\tau$ and if so, the algorithm outputs $\bar{v} = \frac{\widehat{\text{UB}}+\widehat{\text{LB}}}{2}$. This second case is for the setting where $\bar{v}^*(s) = 0$, or very near zero, meaning it would not terminate in Case 1. The relative error will remain high, even though $\bar{v}(s)$ is sufficiently close to $\bar{v}^*(s)$ to satisfy (8) because of $\tau > 0$.

We show that for both cases, (8) is satisfied. We begin with the proof for Case 1. Assume the algorithm terminated, according to the condition in Case 1. For all $j \in \{1, \dots, t\}$, $c_j > 0$ and UB > 0 since UB $=$ $\min_j(|\bar{g}_j| + c_j)$. Upon termination, we have with probability $1-\delta$,

$$\begin{aligned}
|\bar{v}(s)| &= \frac{(1+\epsilon)\text{LB} + (1-\epsilon)\text{UB}}{2} \\
&\le \frac{(1+\epsilon)\text{LB} + (1+\epsilon)\text{LB} + 2\epsilon\tau}{2} \\
&= (1+\epsilon)\text{LB} + \epsilon\tau \\
&\le (1+\epsilon)|\bar{v}^*(s)| + \epsilon\tau.
\end{aligned}$$

Similarly,

$$\begin{aligned}
|\bar{v}(s)| &= \frac{(1+\epsilon)\text{LB} + (1-\epsilon)\text{UB}}{2} \\
&\ge \frac{(1-\epsilon)\text{UB} + (1-\epsilon)\text{UB} + 2\epsilon\tau}{2} \\
&\ge (1-\epsilon)|\bar{v}^*(s)| + \epsilon\tau
\end{aligned}$$

Combining these two inequalities gives

$$\big||\bar{v}(s)| - |\bar{v}^*(s)|\big| \le \epsilon|\bar{v}^*(s)| + \epsilon\tau. \tag{20}$$

When termination occurs under Case, we know LB > 0, and so $|\bar{g}_t| \ge c_t \ge |\bar{g}_t - \bar{v}^*(s)|$. This is because $|\bar{g}_t| - c_t$ must have increased the lower bound, to allow termination. This inequality, $|\bar{g}_t| \ge |\bar{g}_t - \bar{v}^*(s)|$ is only possible if $\bar{v}^*(s)$ is of the same sign as $\bar{g}_t$. This gives that $\text{sign}(\bar{v}(s))$ = $\text{sign}(\bar{g}_t)$ = $\text{sign}(\bar{v}^*(s))$. Because the signs match, $\big||\bar{v}(s)| - |\bar{v}^*(s)|\big| = |\bar{v}(s) - \bar{v}^*(s)|$, and so the result follows from Equation (20).

For Case 2, the interval $[\widehat{\text{LB}}, \widehat{\text{UB}}]$ represents the confidence interval from the IID samples that contains the true mean $\bar{v}^*$. The terminating condition is $\frac{\widehat{\text{UB}}-\widehat{\text{LB}}}{2} \le \epsilon\tau$. For $\bar{v}(s) = \frac{\widehat{\text{UB}}+\widehat{\text{LB}}}{2}$, this gives UB $- \bar{v}(s) = \frac{\widehat{\text{UB}}-\widehat{\text{LB}}}{2} \le \epsilon\tau$ and $\bar{v}(s) -$ LB $= \frac{\widehat{\text{UB}}-\widehat{\text{LB}}}{2} \le \epsilon\tau$. Upon termination, therefore, we have $\epsilon\tau \ge \widehat{\text{UB}} - \bar{v} \ge \bar{v}^* - \bar{v}$ and $\epsilon\tau \ge \bar{v} - \widehat{\text{LB}} \ge \bar{v} - \bar{v}^*$. Thus, $|\bar{v} - \bar{v}^*| \le \epsilon\tau \le \epsilon|\bar{v}^*| + \epsilon\tau$. ■

B HIGH CONFIDENCE BOUNDS FOR CLIPPED MAVE AND MSVE

If one desires to use non-percentage losses, corresponding high-confidence sample complexity bounds are derivable. In this section, we will extend our analysis to the clipped Mean Absolue Value Error (CMAVE) and clipped Mean Squared Value Error (CMSVE).

These are defined as follows:

$$\text{CMAVE}(\hat{v}, \bar{v}) \overset{\text{def}}{=} \mathbb{E}\left[\min(c, |\hat{v}(s_i) - \bar{v}(s_i)|)\right]$$

$$\text{CMSVE}(\hat{v}, \bar{v}) \overset{\text{def}}{=} \mathbb{E}\left[\min(c, (\hat{v}(s_i) - \bar{v}(s_i))^2)\right]$$

Along with their empirical approximations:

$$\text{CMAVE}(\hat{v},\bar{v}) \approx \frac{1}{m}\sum_{i=1}^{m}\min(c,|\hat{v}(s_i)-\bar{v}(s_i)|)$$

$$\text{CMSVE}(\hat{v},\bar{v}) \approx \frac{1}{m}\sum_{i=1}^{m}\min(c,(\hat{v}(s_i)-\bar{v}(s_i))^2)$$

In proving the sample complexity results, we use some of the ideas used in proving Theorem 1. Since both CMAVE and CMSVE are non-percentage losses and do not require a division by the value function, the analysis is greatly simplified. In fact, they no longer require Assumption 2 and, hence, remove the need for EBGStop-like algorithms (1 and 2) presented in Section 4 (which deal with relative errors). Instead of using EBGStop to provide an estimate of the value function, we can simply compute the appropriate number of truncated rollouts (sampled returns) to achieve an estimate of the desired accuracy. These sample complexity numbers are provided in the following analysis.

B.1 SAMPLE COMPLEXITY ANALYSIS OF CLIPPED MAVE

In this section, we will use $\ell(\hat{v},\bar{v})$ to refer CMAVE$(\hat{v},\bar{v})$. Also, the following definitions will be necessary for our analysis:

$$\begin{aligned}\ell_c(\hat{v}(s_i),\bar{v}(s_i)) &\overset{\text{def}}{=} \min(c,|\hat{v}(s_i)-\bar{v}(s_i)|)\\ \hat{\ell}(\hat{v},\bar{v}) &\overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m}\min(c,|\hat{v}(s_i)-\bar{v}(s_i)|)\\ \ell(\hat{v},\bar{v}) &\overset{\text{def}}{=} \mathbb{E}[\hat{\ell}(\hat{v},\bar{v})]\end{aligned}$$

We also define similar quantities replacing $\bar{v}$ with v^* in the above definitions. Below, we present the sample complexity bound for CMAVE.

Theorem 8. *Let $\{s_1,\ldots,s_m\}$ be states sampled I.I.D according to d and that the number of rollouts for each state be n. Let $\bar{\sigma}_i$ be the standard deviation of the rollouts for state i.*

With probability at least $1-\delta$ the following bound for clipped MAVE holds:

$$\left|\ell(\hat{v},v^*)-\hat{\ell}(\hat{v},\bar{v})\right| \le \sqrt{\frac{\log(4K/\delta)c^2}{2m}}+\zeta. \tag{21}$$

for $\zeta = 3R_{max}\left(\frac{1-\gamma^l}{1-\gamma}\right)\frac{\log(6m/\delta)}{n} + \frac{\sum_{i=1}^{m}\bar{\sigma}_i}{m}\sqrt{\frac{2\log(6m/\delta)}{n}} + R_{max}\frac{\gamma^l}{1-\gamma}.$

Proof: Similar to Theorem 1, we start by bounding $\left|\ell(\hat{v},v^*)-\hat{\ell}(\hat{v},\bar{v})\right|$:

$$\begin{aligned}\left|\ell(\hat{v},v^*)-\hat{\ell}(\hat{v},\bar{v})\right| \le{}& \left|\ell(\hat{v},v^*)-\hat{\ell}(\hat{v},v^*)\right|\\ &+\left|\hat{\ell}(\hat{v},v^*)-\hat{\ell}(\hat{v},\bar{v})\right|\end{aligned}$$

The first term is bounded by Hoeffding's inequality in Lemma 2 with probability at least $1-\delta/2$ which gives:

$$\left|\ell(\hat{v},v^*)-\hat{\ell}(\hat{v},v^*)\right| \le \sqrt{\frac{\log(4K/\delta)c^2}{2m}}$$

The second term is bounded in the following way:

$$\begin{aligned}&\left|\hat{\ell}(\hat{v},v^*)-\hat{\ell}(\hat{v},\bar{v})\right|\\ &\le \frac{1}{m}\sum_{i=1}^{m}|\ell_c(\hat{v}(s_i),v^*(s_i))-\ell_c(\hat{v}(s_i),\bar{v}(s_i))|\end{aligned}$$

We can bound each one of these terms as follows:

$$\begin{aligned}&|\ell_c(\hat{v}(s_i),v^*(s_i))-\ell_c(\hat{v}(s_i),\bar{v}(s_i))|\\ &=|\min(c,|\hat{v}(s_i)-v^*(s_i)|)-\min(c,|\hat{v}(s_i)-\bar{v}(s_i)|)|\\ &\le \max\Bigg(\Big|\min\big(c,|\hat{v}(s_i)-\bar{v}(s_i)|+|\bar{v}(s_i)-v^*(s_i)|\big)-\\ &\quad\min\big(c,|\hat{v}(s_i)-\bar{v}(s_i)|\big)\Big|,\Big|\min\big(c,|\hat{v}(s_i)-v^*(s_i)|\big)-\\ &\quad\min\big(c,|\hat{v}(s_i)-v^*(s_i)|+|\bar{v}(s_i)-v^*(s_i)|\big)\Big|\Bigg)\\ &\le|\bar{v}(s_i)-v^*(s_i)|\le|\bar{v}(s_i)-\bar{v}^*(s_i)|+|\bar{v}^*(s_i)-v^*(s_i)|\\ &\le\zeta.\end{aligned} \tag{22}$$

Now, we need to find an expression of ζ. The term $|\bar{v}(s_i)-\bar{v}^*(s_i)|$ can be bounded using the empirical bernstein inequality for random variables with range: $R_{\text{max}}\left(\frac{1-\gamma^l}{1-\gamma}\right)$. This leads us to a bound: $|\bar{v}(s_i)-\bar{v}^*(s_i)| \le 3R_{\text{max}}\left(\frac{1-\gamma^l}{1-\gamma}\right)\frac{\log(6m/\delta)}{n} + \bar{\sigma}_i\sqrt{\frac{2\log(6m/\delta)}{n}}$. The second term can be bounded based on the proof of Lemma 5: $|\bar{v}^*(s_i)-v^*(s_i)| \le R_{\text{max}}\frac{\gamma^l}{1-\gamma}$. This gives us $\zeta = 3R_{\text{max}}\left(\frac{1-\gamma^l}{1-\gamma}\right)\frac{\log(6m/\delta)}{n} + \frac{\sum_{i=1}^{m}\bar{\sigma}_i}{m}\sqrt{\frac{2\log(6m/\delta)}{n}} + R_{\text{max}}\frac{\gamma^l}{1-\gamma}$. We finish the proof by pointing out that due to using hoeffding bound twice with error probability of atmost $\delta/2$ and due to the union bound (to ensure that the bound holds for all m states), the probability that the final bound holds is with at least $1-\delta$. ■

B.2 SAMPLE COMPLEXITY ANALYSIS OF CLIPPED MSVE

In this section, we will use $\ell(\hat{v}, \bar{v})$ to refer CMSVE($\hat{v}, \bar{v}$). Also, the following definitions will be necessary for our analysis:

$$\begin{aligned}
\ell_c(\hat{v}(s_i), \bar{v}(s_i)) &\overset{\text{def}}{=} \min(c, (\hat{v}(s_i) - \bar{v}(s_i))^2) \\
\hat{\ell}(\hat{v}, \bar{v}) &\overset{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \min(c, (\hat{v}(s_i) - \bar{v}(s_i))^2) \\
\ell(\hat{v}, \bar{v}) &\overset{\text{def}}{=} \mathbb{E}[\hat{\ell}(\hat{v}, \bar{v})]
\end{aligned}$$

Similarly, in the above definitions, v^* can be used instead of $\bar{v}$. Below, we present the sample complexity bound for CMSVE.

Theorem 9. *Let $\{s_1, \dots, s_m\}$ be states sampled I.I.D according to d and that the number of rollouts for each state be n. Let $\bar{\sigma}_i$ be the standard deviation of the rollouts for state i.*

With probability at least $1 - \delta$ the following bound for clipped MSVE holds:

$$\left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right| \le \sqrt{\frac{\log(4K/\delta)c^2}{2m}} + \zeta \tag{23}$$

for $\zeta = 3R^2_{max}\left(\frac{1-\gamma^l}{1-\gamma}\right)^2 \frac{\log(6m/\delta)}{n} + \frac{\sum_{i=1}^m \bar{\sigma}_i}{m}\sqrt{\frac{2\log(6m/\delta)}{n}} + R^2_{max}\left(\frac{\gamma^l}{1-\gamma}\right)^2$

Proof: Similar to Theorem 1, we start by bounding $\left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right|$:

$$\begin{aligned}
\left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right| \le{}& \left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, v^*)\right| \\
&+ \left|\hat{\ell}(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right|
\end{aligned}$$

The first term is bounded by Hoeffding's inequality in Lemma 2 with probability atleast $1 - \delta/2$ which gives:

$$\left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, v^*)\right| \le \sqrt{\frac{\log(4K/\delta)c^2}{2m}}$$

The second term is bounded in the following way:

$$\begin{aligned}
&\left|\hat{\ell}(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right| \\
&\le \frac{1}{m}\sum_{i=1}^{m} |\ell_c(\hat{v}(s_i), v^*(s_i)) - \ell_c(\hat{v}(s_i), \bar{v}(s_i))|
\end{aligned}$$

We can bound each one of these terms as follows:

$$\begin{aligned}
&|\ell_c(\hat{v}(s_i), v^*(s_i)) - \ell_c(\hat{v}(s_i), \bar{v}(s_i))| \\
&= \left|\min(c, (\hat{v}(s_i) - v^*(s_i))^2) - \min(c, (\hat{v}(s_i) - \bar{v}(s_i))^2)\right| \\
&\le \max\Bigg(\bigg|\min\left(c, |\hat{v}(s_i) - \bar{v}(s_i)|^2 + |\bar{v}(s_i) - v^*(s_i)|^2\right) - \\
&\min\left(c, |\hat{v}(s_i) - \bar{v}(s_i)|^2\right)\bigg|, \bigg|\min\left(c, |\hat{v}(s_i) - v^*(s_i)|^2\right) - \\
&\min\left(c, |\hat{v}(s_i) - v^*(s_i)|^2 + |\bar{v}(s_i) - v^*(s_i)|^2\right)\bigg|\Bigg) \\
&\le |\bar{v}(s_i) - \bar{v}^*(s_i)|^2 + |\bar{v}^*(s_i) - v^*(s_i)|^2 \le \zeta \qquad (24)
\end{aligned}$$

The first inequality is due to $|a - b|^2 \le (|a - c| + |c - b|)^2$ and this implies $|a - b|^2 \le |a - c|^2 + |c - b|^2 \le (|a - c| + |c - b|)^2$. The range of $\bar{v}$ is $R^2_{\text{max}}\left(\frac{1-\gamma^l}{1-\gamma}\right)^2$. Now, using the empirical bernstein inequality, we can follow the proof technique in Theorem 8 to show that $\zeta = 3R^2_{\text{max}}\left(\frac{1-\gamma^l}{1-\gamma}\right)^2 \frac{\log(6m/\delta)}{n} + \frac{\sum_{i=1}^m \bar{\sigma}_i}{m}\sqrt{\frac{2\log(6m/\delta)}{n}} + R^2_{\text{max}}\left(\frac{\gamma^l}{1-\gamma}\right)^2$.

We finish the proof similarly to conclude that the final bound holds with probability atleast $1 - \delta$ due to the application of hoeffding's bound twice with error probability of atmost $\delta/2$ and due to the union bound. ■

C SAMPLE COMPLEXITY ANALYSIS OF UNCLIPPED LOSSES

Sometimes, one may prefer to consider unclipped losses. Here, we will present sample complexity bounds for the Mean Absolute Value Error (MAVE) and the Mean Squared Value Error (MSVE). To derive meaningful bounds for unbounded random variables, we need to impose other assumptions. There are various options but we choose to explore one: using sub-exponential random variables. With this assumption, the proof techniques mostly follow from those in Section B of the appendix. Below we briefly discuss sub-exponential random variables and illustrate how one can derive the corresponding concentration bounds.

C.1 SUB-EXPONENTIAL CONCENTRATION ANALYSIS

It is well known that for unbounded random variables, finite high probability bounds are not derivable unless it is possible to assume a bound on the moment generating function. One way to derive a meaningful bound is to assume that the tails of the random variable's distribution decay exponentially. If we know the tail decay like a Gaussian distribution, sub-gaussianity is a common assumption. A weaker assumption is sub-exponentiality, which only requires that the moment generating function exists. The Laplace and exponential distributions are two such common fat-tailed distributions. In this section, we will assume that the loss random variable is sub-exponential and derive finite sample complexity bounds. For completeness, we provide the necessary definitions.

Definition C.1. A sub-gaussian random variable X with mean $\mu = \mathbb{E}[X]$ and parameters $\sigma \geq 0$ has the following bound on its moment generating function (MGF):

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}} \qquad \forall \lambda \in \mathbb{R} \tag{25}$$

Definition C.2. A sub-exponential random variable X with mean $\mu = \mathbb{E}[X]$ and parameters $\alpha, \beta \geq 0$ has the following bound on its moment generating function:

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\alpha^2\lambda^2}{2}} \qquad \forall |\lambda| \leq \beta \tag{26}$$

Note that all sub-gaussian RVs are sub-exponential with $\alpha = \sigma$ and $\beta = \infty$, but not all sub-exponential RVs are sub-gaussian. For example, the gaussian distribution is a sub-exponential RV with α being the standard deviation and $\beta = \infty$. Thus, if the loss is known to be sub-gaussian, one can still use the sub-exponential concentration bound. Below, in Theorem 10, we present a concentration bound for sub-exponential random variables.

Theorem 10. *If X_i are I.I.D sub-exponential RVs with parameters (α,β) as defined in Definition C.2, then the following concentration bound holds:*

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right] \leq 2e^{-\frac{nt^2}{2\alpha^2}} \quad \textit{for } \ 0 < t < \alpha^2\beta$$

$$Pr\left[\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| \geq t\right] \leq 2e^{-\frac{nt\beta}{2}} \quad \textit{for } \ t > \alpha^2\beta$$

Proof: For $\lambda \geq 0$,

$$\begin{aligned} Pr\left[\frac{\sum_{i=1}^{n} X_i}{n} - \mu \geq t\right] &= Pr\left[e^{\lambda\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)} \geq e^{\lambda t}\right] \\ &\leq \frac{E\left[e^{\lambda\left(\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right)}\right]}{e^{\lambda t}} \\ &= \frac{E\left[e^{\lambda\left(\frac{\sum_{i=1}^{n}(X_i-\mu)}{n}\right)}\right]}{\prod_{i=1}^{n} e^{\frac{\lambda t}{n}}} \\ &= \prod_{i=1}^{n}\left(\frac{E\left[e^{\lambda\left(\frac{(X_i-\mu)}{n}\right)}\right]}{e^{\frac{\lambda t}{n}}}\right). \end{aligned}$$

We used Markov's inequality and the last equality is due to the independence of X_i. Now, we can bound the moment generating function of each X_i using the sub-exponential RV's property (Definition C.2).

$$\therefore \frac{E[e^{\lambda(X_i-\mu)/n}]}{e^{\lambda t/n}} \leq e^{\frac{0.5\alpha^2\lambda^2 - \lambda t}{n}}$$

Optimizing over λ will result in the tightest bound possible. The minimum of $\frac{0.5\alpha^2\lambda^2-\lambda t}{n}$ is reached at $\lambda = \frac{t}{\alpha^2}$. Replacing λ, we arrive at the expression: $-\frac{t^2}{2\alpha^2}$. By definition $\lambda < \beta$, which results in $t < \alpha^2\beta$. As a result, the following bound for $t \in (0, \alpha^2\beta)$ has to hold:

$$\frac{E[e^{\lambda(X_i-\mu)/n}]}{e^{\lambda t/n}} \leq e^{-\frac{t^2}{2\alpha^2}}$$

$$\therefore \prod_{i=1}^{n}\left(\frac{E\left[e^{\lambda\left(\frac{(X_i-\mu)}{n}\right)}\right]}{e^{\frac{\lambda t}{n}}}\right) \leq \prod_{i=1}^{n} e^{-\frac{t^2}{2\alpha^2}} = e^{-\frac{nt^2}{2\alpha^2}}$$

The same argument follows for the lower tail and by the union bound we conclude that $Pr\left[\left|\frac{\sum_{i=1}^{n} X_i}{n} - \mu\right| \geq t\right] \leq 2e^{-\frac{nt^2}{2\alpha^2}}$. For $t \geq \alpha^2\beta$, the function $0.5\alpha^2\lambda^2 - \lambda t$ decreases monotonously as λ increases since the gradient: $\lambda\alpha^2 - t$ is negative for $0 \leq \lambda < \beta, t \geq \alpha^2\beta$ and thus the minimum is reached at $\lambda = \beta$. So, $0.5\alpha^2\lambda^2 - \lambda t < -\beta t + \frac{\beta^2\alpha^2}{2} \leq -\beta t + \frac{\beta t}{2} = -\frac{\beta t}{2}$. The last inequality is due to $t \geq \alpha^2\beta$. For $t > \alpha^2\beta$, the strict inequality becomes an inequality, resulting in $\frac{E[e^{\lambda(X_i-\mu)/n}]}{e^{\lambda t/n}} \leq e^{-\frac{t\beta}{2}}$. Using the same argument for the confidence bound for the case that $t < \alpha^2\beta$, we conclude the proof.

$\blacksquare$

A key point to notice in Theorem 10 is that sub-exponential variables exhibit gaussian-like tail decay for a small deviation t in contrast to a slower fat tailed decay for larger t. Also, note that a given distribution may be sub-exponential with multiple settings of α and β. To obtain the best concentration bounds, we would want to optimize these parameters, a task which will depend on the exact distribution being considered.

To give an example of how one can prove sub-exponentiality of random variables, we analyze the Laplace distribution.

Definition C.3. (Laplace MGF) If $X \sim$ Lap(μ, b) with probability density function $= \frac{1}{2b}e^{\frac{|x-\mu|}{b}}$, then $\mathbb{E}[e^{\lambda(X-\mu)}] = \frac{1}{1-b^2\lambda^2}$ for $|\lambda| < \frac{1}{b}$

Proposition 11. *If $X \sim$ Lap(μ,b), then X is a sub-exponential RV with $\alpha = b\sqrt{5.12}$ and $\beta = \frac{\sqrt{0.9}}{b}$.*

Proof: Notice that $\frac{1}{1-x} \leq e^{2.56x}$ for $0 \leq x \leq 0.9$. The second inequality comes from basic calculations that conclude $e^{2.55x} \approx \frac{1}{1-x}$ for $x = 0.9$, $e^{2.56x} > \frac{1}{1-x}$, and the fact that $e^{2.56x}$ is always above the function $\frac{1}{1-x}$ for $0 \leq x \leq 0.9$. Based on the above inequality, for $X \sim$ Lap(μ,b), $\mathbb{E}[e^{\lambda(X-\mu)}] = \frac{1}{1-b^2\lambda^2} \leq e^{2.56b^2\lambda^2} = e^{\frac{(\sqrt{5.12}b)^2\lambda^2}{2}}$. Thus, $\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\alpha^2\lambda^2}{2}}$ for $\alpha = b\sqrt{5.12}$ and $|\lambda| < \beta = \frac{\sqrt{0.9}}{b}$. This concludes the proof. ∎

Note that these constants for α and β were not optimized in the above proof and the given values are only one parameter setting out of (infinitely) many that show that the Laplace distribution is sub-exponential.

C.2 SAMPLE COMPLEXITY ANALYSIS OF UNCLIPPED MAVE

In this section we assume that the loss is a sub-exponential random variable. Following the proof of Theorem 8, it is not hard to notice that the only difference will be replacing the Hoeffding's confidence bound with the sub-exponential concentration bound. The following corollary states the result.

Corollary 2. *Let $\{s_1, \ldots, s_m\}$ be states sampled I.I.D according to d and that the number of rollouts for each state be n.*

If the loss is a sub-exponential random variable with parameters α and β, with probability at least $1 - \delta$ the following bound for unclipped MAVE holds:

$$\left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right| \leq t + \zeta. \tag{27}$$

for $\zeta = 3R_{max}\left(\frac{1-\gamma^l}{1-\gamma}\right)\frac{\log(6m/\delta)}{n} + \frac{\sum_{i=1}^m \bar{\sigma}_i}{m}\sqrt{\frac{2\log(6m/\delta)}{n}} + R_{max}\frac{\gamma^l}{1-\gamma}.$

Let $\sigma_1 = \alpha\sqrt{\frac{2\log(4K/\delta)}{m}}$ *and* $\sigma_2 = \frac{2\log(4K/\delta)}{\beta m}$. *If* $0 < \sigma_1 \leq \alpha^2\beta$ *and* $0 < \sigma_2 \leq \alpha^2\beta$, *then* $t = \sigma_1$. *If* $\sigma_1 > \alpha^2\beta$ *and* $\sigma_2 > \alpha^2\beta$, *then* $t = \sigma_2$.

Proof: Let X_i be the empirical loss for each state i and the mean loss be μ. Due to Theorem 10, for K different empirical loss mean estimates and for $0 < \sigma_1 \leq \alpha^2\beta$, setting $Pr\left[\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \geq \sigma_1\right] \leq 2e^{-\frac{m\sigma_1^2}{2\alpha^2}} = \delta/2K$, gives us $\sigma_1 = \alpha\sqrt{\frac{2\log(4K/\delta)}{m}}$.

For $\sigma_2 > \alpha^2\beta$, setting $Pr\left[\left|\frac{1}{m}\sum_{i=1}^m X_i - \mu\right| \geq \sigma_2\right] \leq 2e^{-\frac{m\sigma_2\beta}{2}} = \delta/2K$, gives us $\sigma_2 = \frac{2\log(4K/\delta)}{\beta m}$.

Based on the conditions for σ_1 and σ_2 to be valid, t is chosen accordingly. Thus, using union bound over K empirical loss mean estimates, the total error probability is at most $\delta/2$. The rest of the results regarding ζ follows from Theorem 8 since the bounding technique in its proof does not rely on clipping even though the loss is clipped. This later part gives an error probability of at most $\delta/2$ and so the total error probability is at most δ. ∎

C.3 SAMPLE COMPLEXITY ANALYSIS OF UNCLIPPED MSVE

Corollary 3. *Let $\{s_1, \ldots, s_m\}$ be states sampled I.I.D according to d and that the number of rollouts for each state be n.*

If the loss is a sub-exponential random variable with parameters α and β, with probability at least $1 - \delta$ the following bound for unclipped MSVE holds:

$$\left|\ell(\hat{v}, v^*) - \hat{\ell}(\hat{v}, \bar{v})\right| \leq t + \zeta. \tag{28}$$

for $\zeta = 3R_{max}^2\left(\frac{1-\gamma^l}{1-\gamma}\right)^2\frac{\log(6m/\delta)}{n} + \frac{\sum_{i=1}^m \bar{\sigma}_i}{m}\sqrt{\frac{2\log(6m/\delta)}{n}} + R_{max}^2\left(\frac{\gamma^l}{1-\gamma}\right)^2$

Let $\sigma_1 = \alpha\sqrt{\frac{2\log(4K/\delta)}{m}}$ *and* $\sigma_2 = \frac{2\log(4K/\delta)}{\beta m}$. *If* $0 < \sigma_1 \leq \alpha^2\beta$ *and* $0 < \sigma_2 \leq \alpha^2\beta$, *then* $t = \sigma_1$. *If* $\sigma_1 > \alpha^2\beta$ *and* $\sigma_2 > \alpha^2\beta$, *then* $t = \sigma_2$.

Proof: The same argument from Corollary 2 is applied here for choosing t appropriately. Similarly, the rest of the results regarding ζ follows from Theorem 9 since the bounding technique in its proof does not rely on clipping even though the loss is clipped. ∎

D ALGORITHM DETAILS

In this section, we provide additional details on the pseudocode in the main body, as well as providing the replacement for Algorithm 2 for other the losses discussed in the appendix.

D.1 Sampling returns

To sample the returns to satisfy Assumption 1 for the discounted setting, we provide Algorithm 3. We use the result in Lemma 3 to ensure Assumption 1 is satisfied.

There are a few other details that warrant explanation in the pseudocode for Algorithm 2. The trajectory rollouts

Algorithm 3 Sample truncated return to satisfy Assumption 1

```
▷ Input $\epsilon, \delta, \tau, \gamma$, state $s$
▷ Output a sampled return, $g$
$p_\gamma = 1$
$g \leftarrow 0$
$s_0 \leftarrow s$
while $p_\gamma > \epsilon(1-\gamma)/R_{\max}$ do
    Sample next $s_{k+1}, r_{k+1}$, sampling the action according to $\pi(\cdot|s_k)$
    $g \leftarrow g + p_\gamma r_{k+1}$
    $p_\gamma \leftarrow p_\gamma \gamma$
return $g$
```

Algorithm 4 Empirical confidence interval using bootstrapping

```
▷ Input number of sets to sample $k$ (e.g., k = 1000), and iteration $j$.
$\delta' \leftarrow \frac{3}{(3/d_h)^\alpha} \cdot 100$
$D \leftarrow$ randomly sample $k$ sets of size $j$ from the empirical distribution $\hat{F}$
$\{g_1, \dots, g_k\} \leftarrow$ compute the means from the sets in $D$
$c_{\delta'} \leftarrow$ the $\delta'$'th percentile from $\{g_1, \dots, g_k\}$
$c_{100-\delta'} \leftarrow$ the $(100 - \delta')$'th percentile from $\{g_1, \dots, g_k\}$
$c_j \leftarrow \max(c_{\delta'}, c_{100-\delta'})$
LB $\leftarrow \max(\text{LB}, |\bar{g}| - c_j)$
UB $\leftarrow \min(\text{UB}, |\bar{g}| + c_j)$
```

are of the appropriate lengths given by Lemma 5 to ensure the error due to truncation is sufficiently small. For the empirical Bernstein inequality, we need to estimate the mean and variance of the sample truncated returns. We use a numerically stable approach to compute this sample mean and standard deviation, using Welford's algorithm [Welford, 1962].

D.2 Sampling algorithm for CMAVE, CMSVE, MAVE and MSVE

In this section, we present an incremental sampling algorithm (Algorithm 5) that can be used to sample states with their values and hence guarantee that the high probability errors of sub-exponential MAVE, MSVE and clipped MAVE, MSVE are bounded by a desired preset amount ϵ. This algorithm would be called in Algorithm 1, in place of Algorithm 2. Since a given error can be satisfied with different combinations of m — the number of sampled states—and n—the number of rollouts per state—one option for MSVE and MAVE is to pick m such that the error contributed by the sub-exponential bound is atmost $\alpha^2\beta$ to take advantage of the subgaussian tail decay. Such a choice corresponds to $m = \lceil \frac{2\log(4K/\delta)}{\alpha^2\beta^2} \rceil$. For CMAVE, CMSVE, we suggest fixing m beforehand depending on c, ϵ and how costly it is to sample more rollouts compared to sampling states. We leave other selection criteria for future work.

Algorithm 5 High confidence $\bar{v}$ estimator for clipped losses

```
▷ Input $\epsilon, \delta, m, K, \alpha, \beta$
▷ Compute the values $\bar{v}$ once offline and store for repeated use.
▷ If using CMAVE/MAVE, set $V_{\max} = R_{\max}\left(\frac{1-\gamma^l}{1-\gamma}\right)$
▷ If using CMSVE/MSVE, set $V_{\max} = R_{\max}^2\left(\frac{1-\gamma^l}{1-\gamma}\right)^2$
▷ If using MAVE/MSVE, set $m = \lceil \frac{2\log(4K/\delta)}{\alpha^2\beta^2} \rceil$ and $\zeta \leftarrow \epsilon - \alpha\sqrt{\frac{2\log(4K/\delta)}{m}}$
▷ Else for CMAVE/CMSVE: $\zeta \leftarrow \epsilon - \sqrt{\frac{\log(4K/\delta)c^2}{2m}}$
▷ For states $i = 1, .., m$ initialize:
$\bar{g}_i \leftarrow 0$, $M_i \leftarrow 0$
$j_i \leftarrow 1$, $h_i \leftarrow 0$, $\alpha_i \leftarrow 1$, $x_i \leftarrow 1$
$\beta \leftarrow 1.1$, $p \leftarrow 1.1$, $\zeta \leftarrow \epsilon - \sqrt{\frac{\log(4K/\delta)c^2}{2m}}$
while True do
    for $i \in \{1, .., m\}$ do
        $g_i \leftarrow$ Sampled return from state $i$ of length $l$
        $\Delta_i \leftarrow g_i - \bar{g}_i$
        $\bar{g}_i \leftarrow \bar{g}_i + \frac{\Delta_i}{j_i}$
        $M_i \leftarrow M_i + \Delta_i(g_i - \bar{g}_i)$
        $\sigma_i \leftarrow \sqrt{M_i/j_i}$
        ▷ Compute the confidence interval
        if $j_i \geq \lfloor \beta^{h_i} \rfloor$ then
            $h_i \leftarrow h_i + 1$
            $\alpha_i \leftarrow \lfloor \beta^{h_i} \rfloor / \lfloor \beta^{h_i - 1} \rfloor$
            $x_i \leftarrow -\alpha_i \log \frac{\delta(p-1)}{6mph_i^p}$
            $c_i \leftarrow \sigma_i\sqrt{\frac{2x_i}{j_i}} + \frac{3V_{\max}x_i}{j_i}$
        $j_i = j_i + 1$
    if $\frac{\sum_{i=1}^m c_i}{m} \leq \zeta$ then
        ▷ For all states $i = 1, .., m$ :
        $\bar{v}(i) \leftarrow \bar{g}_i$
        return $\bar{v}$
```

D.3 Computation of optimal intervals

For completeness, we include how we used bootstrapping to compute the intervals to provide a similar stopping rule to EBGStop. The algorithm is the same, except in how the confidence intervals are computed. We first generate a large batch of data, to act as the empirical distribution. We could simply sample sets of size j repeatedly, from the simulator, to get a sense of variability of sample averages. However, we choose to sample a very large batch of data upfront, to reduce the computational burden of the procedure. On each step, a large number k of set of j return samples are drawn, and their sample average computed to obtain the spread of values. Then the percentile corresponding to δ is computed, to provide a high-confidence estimate of a lower and an upper bound on the true values. This approach to computing the true confidence interval is given in Algorithm 4. We sampled an batch of 10^7 returns for each state, to provide the empirical distribution, and set $k = 1000$.

This approach is not a suitable strategy to get high confidence estimates, because it requires a very large number of samples. Rather, we only used this strategy as a comparison, to provide a close approximation to the true confidence intervals, and so obtain best-case sampling numbers. This allowed us to evaluated the impact of the looseness of our bounds, in terms of how many extra samples are generated.