A. Proof of HCGA Guarantees

Theorem 1. *If* $\Pr(b(\theta, M_{safety}, \delta) \leq f(\theta)) \geq 1 - \delta$*, then*

$$\Pr[f(\texttt{alg}(M_{acc})) \geq j] \geq 1 - \delta.$$

Proof. In this proof, we will show that for all θ_c in Θ, $\Pr(f(\texttt{alg}(M_{\text{acc}})) \geq j|\Theta_c = \theta_c) \geq 1 - \delta$, and hence that $\Pr(f(\texttt{alg}(M_{\text{acc}})) \geq j) \geq 1 - \delta$, where $\Theta_c \in \Theta$ is the random variable representing the candidate policy in Algorithm 1. We consider two possible cases: **1)** when $f(\Theta_c) \geq j$ and **2)** when $f(\Theta_c) < j$. In the first case $f(\texttt{alg}(M_{\text{acc}})) \geq j$ always since either $\texttt{alg}(M_{\text{acc}}) = \theta_c$ and by assumption $f(\Theta_c) \geq j$, or $\texttt{alg}(M_{\text{acc}}) = \text{NSF}$ and by definition $f(\text{NSF}) = j$. Hence, $\Pr(f(\texttt{alg}(M_{\text{acc}})) \geq j|\Theta_c = \theta_c) = 1 \geq 1 - \delta$.

Next consider the second case. In this case, we have that for all $\theta_c \in \Theta$ such that $f(\theta_c) < j$:

$$\begin{aligned}
\Pr\left(f(\texttt{alg}(M_{\text{acc}})) \geq j \middle| \Theta_c = \theta_c\right) \overset{\text{(a)}}{=} & \Pr\left(\texttt{alg}(M_{\text{acc}}) = \text{NSF} \middle| \Theta_c = \theta_c\right) \\
\overset{\text{(b)}}{=} & \Pr\left(b(\Theta_c, M_{\text{safety}}, \delta) < j \middle| \Theta_c = \theta_c\right) \\
\overset{\text{(c)}}{\geq} & \Pr\left(b(\Theta_c, M_{\text{safety}}, \delta) \leq f(\theta_c) \middle| \Theta_c = \theta_c\right) \\
= & \Pr\left(b(\theta_c, M_{\text{safety}}, \delta) \leq f(\theta_c) \middle| \Theta_c = \theta_c\right) \\
\overset{\text{(d)}}{=} & \Pr\left(b(\theta_c, M_{\text{safety}}, \delta) \leq f(\theta_c)\right) \\
\overset{\text{(e)}}{\geq} & 1 - \delta,
\end{aligned}$$

where **(a)** follows because when the candidate solution is unsafe (that is, when $f(\Theta_c) < j$), $f(\texttt{alg}(M_{\text{acc}})) \geq j$ if and only if $\texttt{alg}(M_{\text{acc}}) = \text{NSF}$; **(b)** follows from lines 3 and 4 of Algorithm 1, which indicate that $\texttt{alg}(M_{\text{acc}}) = \text{NSF}$ if and only if $b(\theta_c, M_{\text{safety}}, \delta) < j$; **(c)** follows because we are considering the second case, wherein $f(\theta_c) < j$; **(d)** follows because M_{safety} and Θ_c are statistically independent random variables due to Θ_c being computed solely from M_{train}, which is statistically independent of M_{safety}, (that is, for all $M_{1:k} \in \mathcal{M}$, $\Pr(M_{\text{safety}} = M_{1:k}|\Theta_c = \theta_c) = \Pr(M_{\text{safety}} = M_{1:k})$); and **(e)** follows from the assumption in the theorem statement that for all $\theta \in \Theta$, $\Pr(b(\theta, M_{\text{safety}}, \delta) \leq f(\theta)) \geq 1 - \delta$. □

B. Expected Return HCGAs with Control Variates Proofs

For all proofs in this section, recall that $\mathbf{E}[(\text{some expression involving } p_i)|M_i \sim \mu]$ means that p_i are the parameters of MDP M_i (and that therefore p_i itself is random).

Property 1. *For all* $\theta \in \Theta$*, for all* $c \in \mathbb{R}$*,* Z_i *is an unbiased estimator of* $J_\mu(\theta)$*.*

Proof.

$$\begin{aligned}
\mathbf{E}[Z_i|M_i \sim \mu] =& \mathbf{E}\Big[J_{M_i}(\theta) - c\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]\Big)\Big|M_i \sim \mu\Big] \\
=& \mathbf{E}[J_{M_i}(\theta)|M_i \sim \mu] - c(\mathbf{E}[\bar{v}_\theta(p_i)|M_i \sim \mu] - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]) \\
=& \mathbf{E}[J_{M_i}(\theta)|M_i \sim \mu].
\end{aligned}$$

□

Assumption 1 states that the learned control variate varies with its input (that is, that the control variant is not a constant), or, in other words, that the variance is not zero. Formally:

Assumption 1. *For the policy parameterized by* $\theta \in \Theta$*,* $\text{Var}(\bar{v}_\theta(p_i)|M_i \sim \mu) > 0$*.*

Property 2.

$$\underset{c \in \mathbb{R}}{\text{argmin}}\, \text{Var}(Z_i|M_i \sim \mu) = \frac{\mathbf{E}\Big[\Big(J_{M_i}(\theta) - \mathbf{E}[J_{M_k}(\theta)|M_k \sim \mu]\Big)\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]\Big)\Big|M_i \sim \mu\Big]}{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu]\Big)^2\Big|M_{i'} \sim \mu\Big]}.$$

Proof. For brevity, in this proof only, we write $\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]$ as A, and we write $J_{M_i}(\theta)$ as J. All expectations in this proof are given $M_i \sim \mu$ (written out fully only on the first line), or $M_{i'} \sim \mu$ ($M_{i'}$ instead of M_i to disambiguate in equations where there are more than one of these expectations). For example, $\mathbf{E}[\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]|M_i \sim \mu]$ is written as $\mathbf{E}[\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]]$ (or simply as $\mathbf{E}[A]$). Recall from the proof of Property 1 that $\mathbf{E}[A] = 0$, a fact that is exploited in the proof below.

First, we derive an expression for the variance:

$$\begin{aligned}
\mathrm{Var}(Z_i|M_i \sim \mu) =& \mathrm{Var}(J - cA) \\
=& \mathbf{E}[(J-cA)^2] - \mathbf{E}[J-cA]^2 \\
=& \mathbf{E}[J^2] - 2c\mathbf{E}[JA] + c^2\mathbf{E}[A^2] - (\mathbf{E}[J] - c\underbrace{\mathbf{E}[A]}_{=0})^2 \\
=& \mathbf{E}[J^2] - 2c\mathbf{E}[JA] + c^2\mathbf{E}[A^2] - \mathbf{E}[J]^2.
\end{aligned}$$

Minimizing with respect to c by solving for the critical points:

$$\begin{aligned}
0 =& \frac{\partial \,\mathrm{Var}(Z_i)}{\partial c} \\
=& -2\mathbf{E}[JA] + 2c\mathbf{E}[A^2].
\end{aligned}$$

Next, we verify that this critical point is a minimum. Consider the second derivative, $\frac{\partial^2 \,\mathrm{Var}(Z_i)}{\partial c^2} = 2\mathbf{E}[A^2]$. $2\mathbf{E}[A^2]$ is positive if $\mathbb{E}[A^2] \neq 0$. $\mathbf{E}[A^2] = \mathbf{E}[(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu])^2] = \mathrm{Var}(\bar{v}_\theta(p_i)|M_i \sim \mu)$. By Assumption 1, $\mathbf{E}[A^2] \neq 0$, so $\mathbf{E}[A^2]$ is positive. Therefore, this critical point is a minimum. Solving for c:

$$\begin{aligned}
c =& \frac{\mathbf{E}[JA]}{\mathbf{E}[A^2]} \\
=& \frac{\mathbf{E}[J(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu])]}{\mathbf{E}[(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu])^2]}.
\end{aligned}$$

Consider the numerator of this fraction (for readability, we stop writing all given terms for the remainder of the proof):

$$\begin{aligned}
\mathbf{E}\Big[J\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)]\Big)\Big] =& \mathbf{E}[J\bar{v}_\theta(p_i)] - \mathbf{E}\Big[J\mathbf{E}[\bar{v}_\theta(p_j)]\Big] \\
\overset{\text{(a)}}{=}& \mathbf{E}[J\bar{v}_\theta(p_i)] - \mathbf{E}[J]\mathbf{E}[\bar{v}_\theta(p_j)] \\
=& \mathrm{Cov}(J, v_\theta(p_i)),
\end{aligned}$$

where **(a)** results from the fact that the expectation of J is with respect to M_i, and that M_i and M_j are independent.

The covariance written as $\mathbf{E}[J(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu])]$ is correct but may be numerically unstable, and so it should not be computed in this form. An equivalent and more numerically stable form is:

$$\begin{aligned}
\mathbf{E}[J(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu])] =& \mathrm{Cov}(J, v_\theta(p_i)) \\
=& \mathbf{E}\Big[\Big(J - \mathbf{E}[J]\Big)\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)]\Big)\Big].
\end{aligned}$$

So,

$$c = \frac{\mathbf{E}\Big[\Big(J - \mathbf{E}[J]\Big)\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)]\Big)\Big]}{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu]\Big)^2\Big]}.$$

□

Corollary 1. *If $\bar{v}_\theta(p_i) = J_{M_i}(\theta)$, then*

$$\operatorname*{argmin}_{c \in \mathbb{R}} \mathrm{Var}(Z_i|M_i \sim \mu) = 1.$$

Proof. By Property 2,

$$c = \frac{\mathbf{E}\Big[\Big(J_{M_i}(\theta) - \mathbf{E}[J_{M_k}(\theta)|M_k \sim \mu]\Big)\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]\Big)\Big|M_i \sim \mu\Big]}{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu]\Big)^2\Big|M_{i'} \sim \mu\Big]}.$$

Substituting the control variate for the objective:

$$\begin{aligned} c =& \frac{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_k)|M_k \sim \mu]\Big)\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]\Big)\Big|M_i \sim \mu\Big]}{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu]\Big)^2\Big|M_{i'} \sim \mu\Big]} \\ =& \frac{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]\Big)^2\Big|M_i \sim \mu\Big]}{\mathbf{E}\Big[\Big(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu]\Big)^2\Big|M_{i'} \sim \mu\Big]} \\ =& 1. \end{aligned}$$

□

C. Expected Return HCGAs with Control Variates

Algorithm 2 Expected Return HCGA with Control Variate Template

```
Input  : Feasible set $\Theta$, a set of MDPs $M_{\text{acc}}$, user-defined threshold $j$, probability $1-\delta$, and high-confidence bounding
         function $b$.
Output : $\theta \in \Theta \cup \{\text{NSF}\}$
Partition $M_{\text{acc}}$ into two data sets, $M_{\text{train}}$ and $M_{\text{safety}}$;
Compute a $\theta_c \in \text{argmax}_{\theta \in \Theta} J_{M_{\text{train}}}(\theta)$;
For all $M_i \in M_{\text{train}}$, compute $J_{M_i}(\theta_c)$;
Use the training data collected above (that is, for all $M_i \in M_{\text{train}}$, $p_i$ and $J_{M_i}(\theta_c)$) to compute some $\bar{v}_{\theta_c}$ (this is a regression
  problem).
Ensure that $\bar{v}_{\theta_c}$ is not a constant (if it is a constant, choose a better function approximator, training or optimization algorithm,
  and/or control variate hyperparameters; alternatively, use a standard HCGA without a control variate).
Using the whole distribution of MDP parameters from $\mu$, estimate (or calculate exactly if possible) $\mathbf{E}[\bar{v}_{\theta_c}(p_j)|M_j \sim \mu]$. For
  brevity, define $e_v$ to be the estimate of this expectation: $e_v := \mathbf{E}[\bar{v}_{\theta_c}(p_j)|M_j \sim \mu]$;
Estimate an optimal $c$ value: Use the training data to estimate $\mathbf{E}\Big[\Big(J_{M_i}(\theta_c) - \mathbf{E}[J_{M_k}(\theta_c)|M_k \sim \mu]\Big)\Big(\bar{v}_{\theta_c}(p_i) - e_v\Big)\Big|M_i \sim \mu\Big]$
  and $\mathbf{E}[(\bar{v}_{\theta_c}(p_{i'}) - e_v)^2|M_{i'} \sim \mu]$, using $J_{\text{train}}(\theta_c)$ to estimate $\mathbf{E}[J_{M_k}(\theta_c)|M_k \sim \mu]$. Use these values (they are the numerator
  and denominator of the following expression) to estimate

      $c = \frac{\mathbf{E}\big[\big(J_{M_i}(\theta_c) - \mathbf{E}[J_{M_k}(\theta_c)|M_k \sim \mu]\big)\big(\bar{v}_{\theta_c}(p_i) - \mathbf{E}[\bar{v}_{\theta_c}(p_j)|M_j \sim \mu]\big)\big|M_i \sim \mu\big]}{\mathbf{E}\big[\big(\bar{v}_{\theta_c}(p_{i'}) - \mathbf{E}[\bar{v}_{\theta_c}(p_{j'})|M_{j'} \sim \mu]\big)^2\big|M_{i'} \sim \mu\big]}.$

  Alternatively, set $c = 1$;
Define $Z_{i''} := J_{M_{i''}}(\theta_c) - c(\bar{v}_{\theta_c}(p_{i''}) - e_v)$. In the bound computation in the next step, for MDPs $M_1, M_2, \dots, M_k$ in
  $M_{\text{safety}}$, use $Z_1, Z_2, \dots, Z_k$ instead of $J_{M_1}(\theta_c), J_{M_2}(\theta_c), \dots, J_{M_k}(\theta_c)$ to compute $J_{M_{\text{safety}}}(\theta_c)$, $\hat{\sigma}_J(\theta_c, M_{\text{safety}})$, and/or any
  other relevant statistics;
if $b(\theta_c, M_{\text{safety}}, \delta) \geq j$ then return $\theta_c$;
else return NSF;
```

Remark: it may be possible to calculate $\mathbf{E}[\bar{v}_{\theta_c}(p_j)|M_j \sim \mu]$ (e_v in the algorithm above) exactly instead of estimating it. For example, if there are finite MDPs in the support of the distribution, and the distribution is uniform over those MDPs, then it

HCGA	Safety Function and Bounding Function	Intuition
Hoeffding	$f(\theta) \coloneqq J_\mu(\theta).$ $b(\theta, M_{\text{safety}}, \delta) \coloneqq J_{M_{\text{safety}}}(\theta) - \sqrt{\ln(1/\delta)/(2\lvert M_{\text{safety}}\rvert)}.$	Safety constraint on the objective.
t-test	$f(\theta) \coloneqq J_\mu(\theta).$ $b(\theta, M_{\text{safety}}, \delta) \coloneqq J_{M_{\text{safety}}}(\theta) - \frac{\hat{\sigma}_J(\theta, M_{\text{safety}})\mathtt{t}_{1-\delta, \lvert M_{\text{safety}}\rvert - 1}}{\sqrt{\lvert M_{\text{safety}}\rvert}}.$	Safety constraint on the objective.
CVaR MDP	$f(\theta) \coloneqq \text{CVaR}_\alpha(J_{M_1}(\theta) \vert M_1 \sim \mu).$ $b(\mathtt{alg}(M_{\text{acc}}), M_{\text{safety}}, \delta)$ $\coloneqq \frac{1}{\alpha}\sum_{i=1}^n (J_{n+1-i} - J_{n-i}) \max(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)).$	Safety constraint on the "worst-case MDPs" in μ. This type of HCGA may be useful if a few rare MDPs in supp (μ) are suspected to entail catastrophic risks. These HCGAs may also be useful when attempting to transfer a policy to a distribution of MDPs, μ', that is similar to μ (that is, the extrapolation setting). Suppose that, due to the similarity between μ and μ', one can reasonably assume that the performance of the policy for the new setting, $J_{\mu'}(\theta)$, will be no worse than the performance for, e.g., the worst 1% of MDPs sampled from μ. Under this type of assumption, a CVaR MDP HCGA can be straightforwardly applied to inform the user whether the policy is likely to achieve safe performance for μ'.
CVaR Episodic	$f(\theta) \coloneqq \text{CVaR}_\alpha(G_{M_1}(\theta) \vert M_1 \sim \mu).$ $b(\mathtt{alg}(M_{\text{acc}}), M_{\text{safety}}, \delta)$ $\coloneqq \frac{1}{\alpha}\sum_{i=1}^n (G_{n+1-i} - G_{n-i}) \max(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)).$	Safety constraint on the worst-case episodes. This type of HCGA may be useful if rare episodes may entail catastrophic risk (e.g., the diabetes setting discussed in Section 8.3).

Figure 2. A summary of the four HCGAs that we study in this work.

may be trivial to calculate e_v exactly by computing $\bar{v}_\theta(p_j)$ for every MDP M_j, and taking the mean of the resulting control variate values.

D. HCGA Summary Table

In Figure 2, we provide a summary of the four HCGAs we study in this paper.

E. Background: VaR and CVaR

Value at risk (VaR) is a measure of risk originally developed as a financial metric to quantify how poorly some set of investments might perform, excluding some proportion of worst-case scenarios. Intuitively, for some random variable X and some proportion α, VaR is simply the α-quantile of X. Formally, for some random variable X and some proportion α, we define VaR as:

$$\text{VaR}_\alpha(X) \coloneqq \inf\{x \in \mathbb{R} \vert \Pr(X \leq x) \geq \alpha\}.$$

Some criticize VaR for being insensitive to catastrophic risks, since it ignores the worst possible outcomes (Brown, 2007).

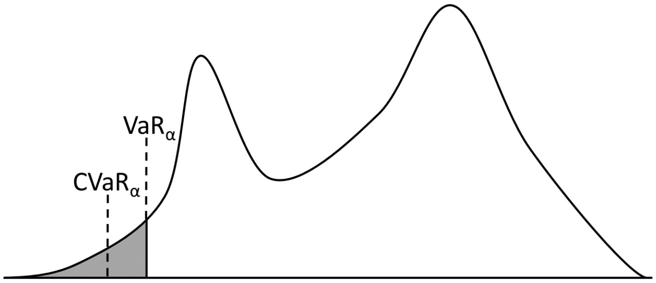


Figure 3. The probability density function of some continuous random variable X. The shaded region has area α. VaR$_\alpha$ is the smallest value such that $\alpha(100\%)$ of samples will be less than it. CVaR$_\alpha$ is the expected value of samples less than or equal to VaR$_\alpha$ (that is, the expected value of samples in the shaded region).

One solution for problems where VaR may not be a suitable measure of risk is *conditional value at risk* (CVaR). Intuitively, given some random variable X and some proportion α, CVaR is the expected value of the lowest α proportion of values of X. In other words, it is the expected value of the "tail" that VaR ignores. Formally, for some continuous random variable X and some proportion α, we define CVaR as:

$$\text{CVaR}_\alpha(X) \coloneqq \mathbf{E}[X|X \leq \text{VaR}_\alpha(X)].$$

For an illustration of VaR and CVaR, see Figure 3.

While this paper restricts itself to CVaR-based HCGAs, one could design VaR-based HCGAs using a high-confidence bound on VaR. Intuitively, VaR-based HCGAs may be appropriate when one wants to ensure that some policy is safe for the majority ($(1-\alpha)100\%$) of MDPs (Section 8.2) or episodes (Section 8.3), but rare, potentially catastrophic risks are either acceptable or nonexistent. CVaR-based HCGAs may be appropriate when one cares less about the overall objective as a measure of safety, but wants to avoid rare catastrophic risks.

F. Brown's CVaR Bound

Let $\hat{C}$ denote the sample-based estimate of CVaR$_\alpha(X)$: $\hat{C} \coloneqq x_{\lceil n\alpha \rceil} - \frac{1}{n\alpha}\sum_{i=1}^{\lceil n\alpha \rceil}(x_{\lceil n\alpha \rceil} - x_i)$, where n is the sample size, and $x_1, ..., x_n$ are the order statistics of the sample (that is, the sample sorted into increasing order). This formulation is equivalent to that of Brown (2007); see supplementary material Section I for the derivation. Brown (2007) bounds CVaR with high confidence:

Property 4. *For all* $\delta \in (0,1)$*, if* $\text{supp}\,(X) \subseteq [a,b]$*:* $\Pr\left(\text{CVaR}_\alpha(X) \geq \hat{C} - (b-a)\sqrt{\frac{5\ln(3/\delta)}{\alpha n}}\right) \geq 1-\delta$*, where* n *is the number of samples of* X *used to calculate* $\hat{C}$*.*

G. Left-Tail Version of Brown's CVaR Bound

Below, we denote the left-tail CVaR that we use as CVaR$^L_\alpha$, and the right-tail CVaR that Brown (2007) used as CVaR$^R_\alpha$. More formally, for a random variable X, we define left and right CVaR respectively, as:

$$\text{CVaR}^L_\alpha(X) \coloneqq \mathbf{E}[X|X \leq \text{VaR}^L_\alpha(X)]$$

and

$$\text{CVaR}^R_\alpha(X) \coloneqq \mathbf{E}[X|X \geq \text{VaR}^R_\alpha(X)],$$

where

$$\text{VaR}^L_\alpha(X) \coloneqq \inf\{x \in \mathbb{R}|\Pr(X \leq x) \geq \alpha\}$$

and

$$\text{VaR}^R_\alpha(X) \coloneqq \sup\{x \in \mathbb{R}|\Pr(X \geq x) \geq \alpha\}.$$

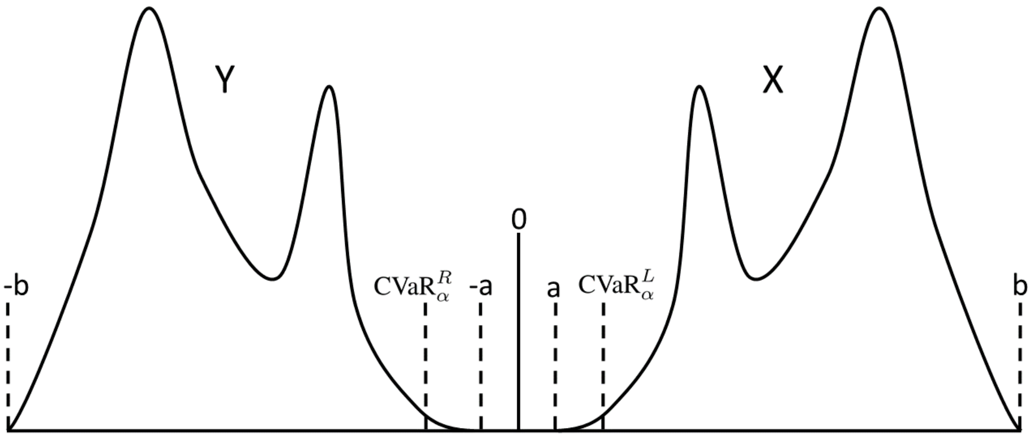


Figure 4. A visualization of the intuition behind the proof of Property 5.

Let $X_1, ...X_n$ be n i.i.d. samples of some continuous random variable X. We denote sample-based estimates of the left and right-tail CVaR values as

$$\hat{C}^L \coloneqq \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{n\alpha} \sum_{i=1}^{n} \max(0, x - X_i) \right\}$$

and

$$\hat{C}^R \coloneqq \inf_{x \in \mathbb{R}} \left\{ x + \frac{1}{n\alpha} \sum_{i=1}^{n} \max(0, X_i - x) \right\},$$

respectively.

In the proof below, we define another continuous random variable Y, such that $Y \coloneqq -X$. See Figure 4 for an intuitive visualization of this setting. We then these two variables to show that the left- and right-tail bounds are equivalent.

Property 5. *For all $\delta \in (0, 1)$, if $\mathrm{supp}(X) \subseteq [a, b]$:*

$$\Pr\left(\mathrm{CVaR}^L_\alpha(X) \geq \hat{C}^L - (b-a)\sqrt{\frac{5\ln(3/\delta)}{\alpha n}} \right) \geq 1 - \delta.$$

Proof. Let $Y \coloneqq -X$. Notice that $\mathrm{supp}(Y) \subseteq [-b, -a]$. First, we show that $\mathrm{CVaR}^R_\alpha(Y) = -\mathrm{CVaR}^L_\alpha(X)$:

$$\begin{aligned}
\mathrm{CVaR}^R_\alpha(Y) =&\mathbf{E}[Y | Y \geq \mathrm{VaR}^R_\alpha(Y)] \\
=&\mathbf{E}[Y | Y \geq \sup\{x \in \mathbb{R} | \Pr(Y \geq x) \geq \alpha\}] \\
=&\mathbf{E}[-X | -X \geq \sup\{x \in \mathbb{R} | \Pr(-X \geq x) \geq \alpha\}] \\
=&\mathbf{E}[-X | X \leq -\sup\{x \in \mathbb{R} | \Pr(-X \geq x) \geq \alpha\}] \\
=&-\mathbf{E}[X | X \leq -\sup\{x \in \mathbb{R} | \Pr(-X \geq x) \geq \alpha\}] \\
=&-\mathbf{E}[X | X \leq -\sup\{x \in \mathbb{R} | \Pr(X \leq -x) \geq \alpha\}] \\
=&-\mathbf{E}[X | X \leq \inf\{-x \in \mathbb{R} | \Pr(X \leq -x) \geq \alpha\}] \\
=&-\mathbf{E}[X | X \leq \inf\{x \in \mathbb{R} | \Pr(X \leq x) \geq \alpha\}].
\end{aligned}$$

Applying the left-tail definitions, we get that

$$\begin{aligned}\text{CVaR}_\alpha^R(Y) =& -\mathbf{E}[X|X \leq \text{VaR}_\alpha^L(X)]\\ =& -\text{CVaR}_\alpha^L(X).\end{aligned}$$

Therefore

$$\text{CVaR}_\alpha^R(Y) = -\text{CVaR}_\alpha^L(X). \tag{4}$$

Next, we show that $\hat{C}_X^L = -\hat{C}_Y^R$. Let $X_1, \dots, X_n$ be n i.i.d. samples of X, and $Y_1, \dots, Y_n$ be n i.i.d. samples of Y, such that $Y_1 \coloneqq -X_1, ..., Y_n \coloneqq -X_n$.

$$\begin{aligned}\hat{C}_X^L \coloneqq& \sup_{x\in\mathbb{R}} \left\{ x - \frac{1}{n\alpha}\sum_{i=1}^n \max(0, x - X_i) \right\}\\ =& \sup_{x\in\mathbb{R}} \left\{ x - \frac{1}{n\alpha}\sum_{i=1}^n \max(0, x + Y_i) \right\}\\ =& -\inf_{x\in\mathbb{R}} \left\{ -x + \frac{1}{n\alpha}\sum_{i=1}^n \max(0, x + Y_i) \right\}\\ =& -\inf_{y\in\mathbb{R}} \left\{ y + \frac{1}{n\alpha}\sum_{i=1}^n \max(0, Y_i - y) \right\}\\ =& -\hat{C}_Y^R.\end{aligned}$$

So

$$\hat{C}_X^L = -\hat{C}_Y^R. \tag{5}$$

Finally, we start with Brown's (2007) right-tail bound for Y:

$$\Pr\left(\text{CVaR}_\alpha^R(Y) \leq \hat{C}_Y^R + ((-a) - (-b))\sqrt{\frac{5\ln(3/\delta)}{\alpha n}} \right) \geq 1 - \delta.$$

Simplifying and applying Equations (4) and (5):

$$\Pr\left(-\text{CVaR}_\alpha^L(X) \leq -\hat{C}_X^L + (b-a)\sqrt{\frac{5\ln(3/\delta)}{\alpha n}} \right) \geq 1 - \delta.$$
$$\Pr\left(\text{CVaR}_\alpha^L(X) \geq \hat{C}_X^L - (b-a)\sqrt{\frac{5\ln(3/\delta)}{\alpha n}} \right) \geq 1 - \delta.$$

□

H. Left-Tail Version of Thomas & Learned-Miller's CVaR Bound

In this section, we prove that the left- and right-tail bounds of Thomas & Learned-Miller (2019) are equivalent. Let $X_1, \dots, X_n$ be n i.i.d. samples of some continuous random variable X, with $\text{supp}(X) \subseteq [a, \infty)$. Let $W_0 \coloneqq a$, and $W_1, \dots, W_n$ be the order statistics of the sample (that is, $X_1, \dots, X_n$ sorted in increasing order). We define $\text{CVaR}_\alpha^L(X)$ and $\text{CVaR}_\alpha^R(X)$ as in Section G above. As in the section above, we define another continuous random variable Y, such that $Y \coloneqq -X$.

Property 6. *For all $\delta \in (0, .5]$:*

$$\Pr\left(\text{CVaR}_\alpha^L(X) \geq W_0 + \frac{1}{\alpha}\sum_{i=1}^{n}(W_{n+1-i} - W_{n-i})\max\left(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)\right) \geq 1 - \delta.$$

Proof. Notice that, since $Y \coloneqq -X$, $\text{supp}(Y) \subseteq (-\infty, -a]$. Let $Y_1, \dots, Y_n$ be n i.i.d. samples of Y, such that $Y_1 \coloneqq -X_1, \dots, Y_n \coloneqq -X_n$. Let $Z_1, \dots, Z_n$ be the order statistics of the sample of Y (that is, $Y_1, \dots, Y_n$ sorted in increasing order), and let $Z_{n+1} \coloneqq -a$.

Theorem 3 of Thomas & Learned-Miller (2019) states that

$$\Pr\left(\text{CVaR}_\alpha^R(Y) \leq Z_{n+1} - \frac{1}{\alpha}\sum_{i=1}^{n}(Z_{i+1} - Z_i)\max\left(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)\right) \geq 1 - \delta.$$

In the proof of Property 5 above, we showed that $\text{CVaR}_\alpha^R(Y) = -\text{CVaR}_\alpha^L(X)$. So

$$\Pr\left(-\text{CVaR}_\alpha^L(X) \leq Z_{n+1} - \frac{1}{\alpha}\sum_{i=1}^{n}(Z_{i+1} - Z_i)\max\left(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)\right) \geq 1 - \delta.$$

Notice that $Z_{n+1} = -a = -W_0$, $Z_n = -W_1, \dots, Z_2 = -W_{n-1}$, $Z_1 = -W_n$. That is, for $j \in \{1, 2, \dots, n, n+1\}$, $Z_j = -W_{n+1-j}$.

Applying these equalities:

$$\Pr\left(-\text{CVaR}_\alpha^L(X) \leq -W_0 - \frac{1}{\alpha}\sum_{i=1}^{n}(-W_{n-i} + W_{n+1-i})\max\left(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)\right) \geq 1 - \delta,$$

so,

$$\Pr\left(\text{CVaR}_\alpha^L(X) \geq W_0 + \frac{1}{\alpha}\sum_{i=1}^{n}(W_{n+1-i} - W_{n-i})\max\left(0, \frac{i}{n} - \sqrt{\frac{\ln(1/\delta)}{2n}} - (1-\alpha)\right)\right) \geq 1 - \delta.$$

□

I. CVaR Estimator Simplification

For some random variable X, given some level $\alpha \in (0, 1)$ and a sample of X of size n, where $x_1, ..., x_n$ are the order statistics of the sample, let $\hat{C}$ denote the sample-based estimate of $\text{CVaR}_\alpha(X)$. We use a definition of $\hat{C}$ that is different from but equivalent to that of Brown (2007); our definition may be more straightforward to implement. We prove that these two definitions are equivalent:

Property 7.

$$\sup_{x\in\mathbb{R}}\left\{x - \frac{1}{n\alpha}\sum_{i=1}^{n}\max(0, x - x_i)\right\} = x_{\lceil n\alpha\rceil} - \frac{1}{n\alpha}\sum_{i=1}^{\lceil n\alpha\rceil}\left(x_{\lceil n\alpha\rceil} - x_i\right).$$

Proof.

$$\begin{aligned}\hat{C} \coloneqq & \sup_{x \in \mathbb{R}} \left\{ x - \frac{1}{n\alpha} \sum_{i=1}^{n} \max(0, x - x_i) \right\} \\ \overset{\text{(a)}}{=} & x_{\lceil n\alpha \rceil} - \frac{1}{n\alpha} \sum_{i=1}^{n} \max(0, x_{\lceil n\alpha \rceil} - x_i) \\ \overset{\text{(b)}}{=} & x_{\lceil n\alpha \rceil} - \frac{1}{n\alpha} \sum_{i=1}^{\lceil n\alpha \rceil} \left(x_{\lceil n\alpha \rceil} - x_i \right),\end{aligned}$$

where **a)** follows from the first two steps of the proof of Proposition 4.1 (Brown, 2007) (see below for their reasoning), and **b)** follows from the fact that the order statistics are non-decreasing and $x_{\lceil n\alpha \rceil} - x_i = 0$ for $i = \lceil n\alpha \rceil$. □

A brief elaboration of the reasoning of Brown (2007) for **(a)**: define

$$g(x) \coloneqq x - \frac{1}{n\alpha} \sum_{i=1}^{n} \max(0, x - x_i).$$

Define

$$h(x, i) \coloneqq \begin{cases} 0, & \text{if } x_i > x; \\ 1, & \text{if } x_i < x; \\ \text{undefined}, & \text{if } x_i = x. \end{cases}$$

Taking the derivative of g, for all $x \notin \{x_1, \dots, x_n\}$:

$$\frac{dg(x)}{dx} = 1 - \frac{1}{n\alpha} \sum_{i=1}^{n} h(x, i).$$

Notice that g is continuous and that, except for the n removable discontinuities, $\frac{dg(x)}{dx}$ is monotonically decreasing (including "across" the discontinuities). Therefore, g is concave. Furthermore, notice that for all $x < x_1$, $\frac{dg(x)}{dx} = 1$, and for all $x > x_n$, $\frac{dg(x)}{dx} = 1 - 1/\alpha$, which is negative for $\alpha \in (0, 1)$. More concisely, the derivative switches signs from positive to negative as x increases.

Therefore, $\sup_{x \in \mathbb{R}} g(x)$ will occur either **1)** when $\frac{dg(x)}{dx} = 0$ (or at points at which the left or right derivative is 0, see Figure 5) or **2)** if for all $x \in \mathbb{R}$, $\frac{dg(x)}{dx} \neq 0$, at the removable discontinuity when $\frac{dg(x)}{dx}$ switches from positive to negative (as in Figure 6). By inspection, the point $x = x_{\lceil n\alpha \rceil}$ is the unique $x \in \mathbb{R}$ that satisfies the criteria in both cases.

J. Environment Descriptions

Generalization gridworld is a 5×5 gridworld with deterministic transitions. The reward is -1 at every time step, except for when the agent is in the terminal state, in which case the reward is 0. Each MDP has "cliff" squares which, if entered, send the agent back to the starting position. A single path from the start state to the goal state is clear of cliffs in all MDPs. Specifically, the following sequence of actions is optimal for all MDPs: RIGHT, DOWN, RIGHT, DOWN, RIGHT, DOWN, DOWN, RIGHT. The result is that, while individual MDPs may have many optimal policies, there is only one optimal policy for the entire set of MDPs. The range of possible returns is $[-200, -7]$.

The dynamics and objective of dynamic arm simulator (DAS1) are fully described by Blana et al. (2009). The arm consists of six muscles and two joints. Episodes are of fixed length, and the reward is proportional to the negative square of the distance between the goal and the endpoint of the arm, with a slight penalty proportional to muscle activation. For DAS1, we make the arm and goal initial state in each MDP deterministic and separate possible initial states into 70^4 MDPs (70 possible values of four angles, two of which describe the arm's starting position and two of which describe the goal). We clip the reward at each time step to be in the interval $[-6, 0]$, so that the normalization of the objective function to the range $[0, 1]$ is easier (rewards less than -6 are quite rare, so this does not have much effect).

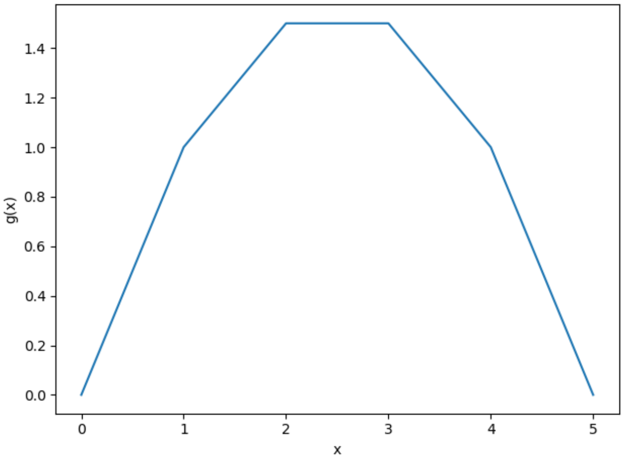


Figure 5. An example of a $g(x)$ for which there exists an x such that $\frac{dg(x)}{dx} = 0$. The supremum is $g(x)$ for all x such that the left and/or right derivatives are 0.

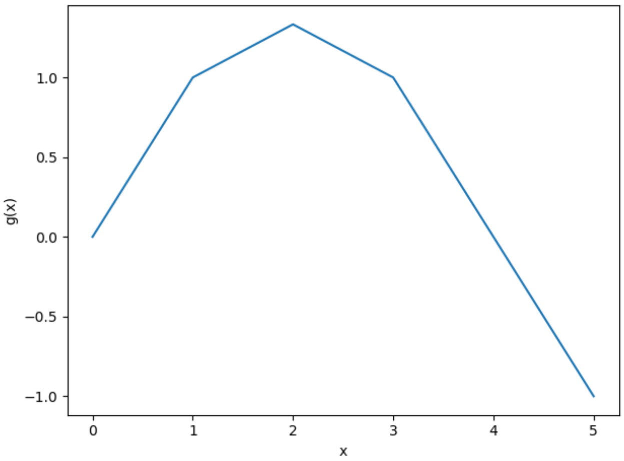


Figure 6. An example of a $g(x)$ for which there does not exist an x such that $\frac{dg(x)}{dx} = 0$. In this case, the supremum lies at the point where $\frac{dg(x)}{dx}$ switches from positive to negative.

K. Full Results and Experimental Details

K.1. Experimental Details

First, we list and discuss the δ, α (for the CVaR quantile, not to be confused with stepsize), and j values used in each experiment. In all experiments, $\delta = 0.1$ and $\alpha = 0.2$.

For the gridworld experiments, $j = -10$ for the Hoeffding HCGA, $j = -8$ for the t-test HCGA (slightly higher than the Hoeffding experiments to highlight the failure behavior with low numbers of MDPs), and $j = -30$ for the CVaR HCGAs. Notice that the CVaR j definitions are significantly lower, since they are for the worst-case tails of the distributions. In Figure 1a above, the standard RL algorithm is plotted using the Hoeffding value, $j = -10$ (the j value affects the plot of the proportion of trials for which the standard RL algorithm failed). A plot of the standard RL algorithm with the t-test value ($j = -8$) is shown in Figure 8b.

For the DAS1 experiments, the Hoeffding and t-test experiments use $j = -25$, and the CVaR experiments use $j = -60$.

In all plots of average returns above, we plot trials which return NSF as j. This choice is because we defined $f(j) \coloneqq \text{NSF}$ and because, intuitively, we have defined NSF to be safe and j is the minimum definition of safe in each experiment. In the plots in Section K.3 below, we provide an alternative interpretation of the same data: excluding NSF trials rather than plotting them as j.

There are four phases to each experiment: **1)** the training phase, in which the candidate policy is trained; **2)** the training evaluation phase, in which the candidate policy's performance is evaluated on the training MDPs; **3)** the safety test phase, in which the policy is run on the safety MDPs, and the safety is test applied; and **4)** the testing phase, in which the candidate policy is run and evaluated on some test set of MDPs. There are always 10,000 test MDPs. For the results to be valid, it is important that sufficient numbers of episodes are run for the training evaluation, safety test, and testing phases.

The number of episodes used for each phase follows. For generalization gridworld, we ran 1024 training episodes per MDP. For DAS1, we ran 10,000 training episodes per MDP.

For the training phase, all MDPs are shuffled into a random order, and each is run once in that order. This process repeats until the maximum number of episodes is run.

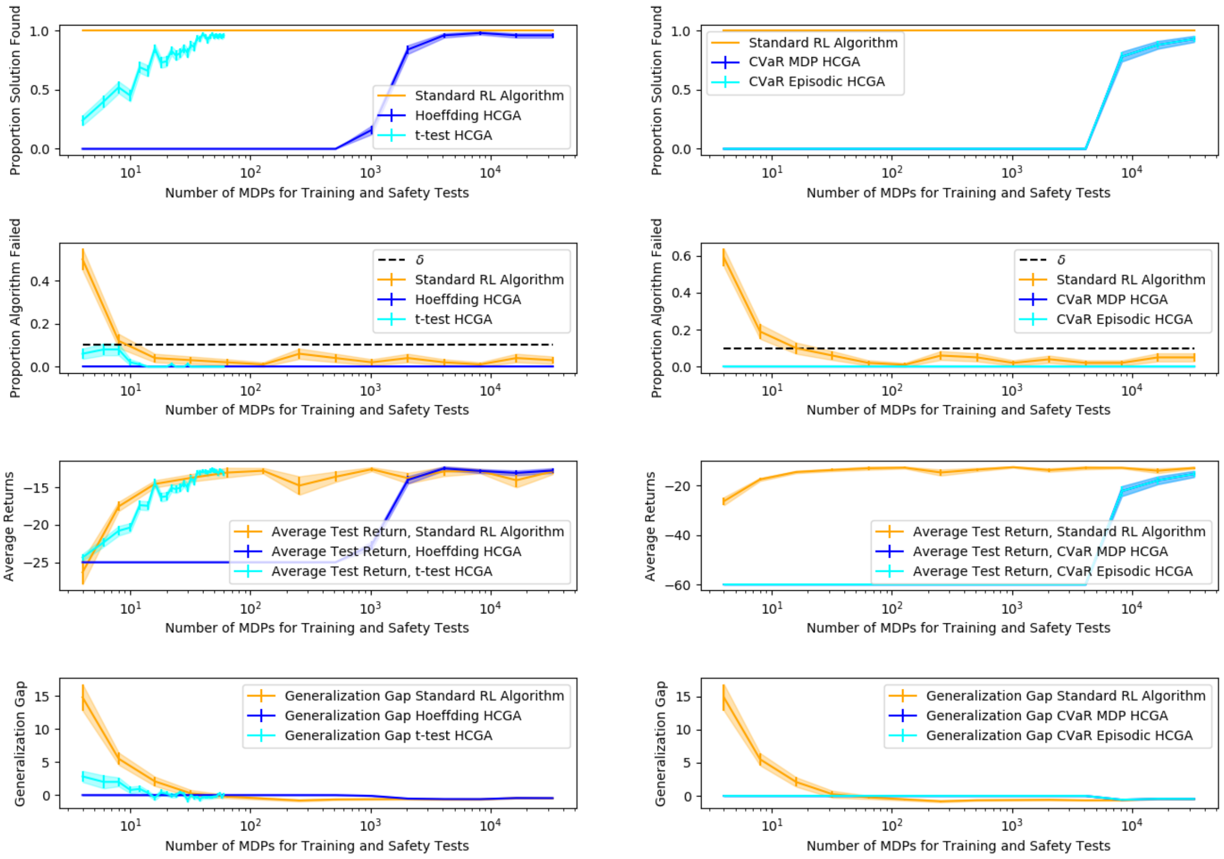
For all experiments, the number of episodes per MDP run in the training evaluation, safety test, and testing phases was $\lceil 10{,}000/n \rceil$, where n is the number of MDPs used in the phase (that is, n is 10,000 for the testing phase, $|M_{\text{train}}|$ for the training evaluation phase, and $|M_{\text{safety}}|$ for the safety testing phase). Notice that, for $n \ll 10{,}000$, this results in approximately 10,000 episodes in the phase. For larger n, this formula also ensures that at least one episode is run for each MDP.

The episodic CVaR HCGA is an exception to the above rule: in the safety test phase, each MDP is only run for one episode, since the safety test samples episodic returns. Sampling a return from each MDP more than once would result in samples not drawn i.i.d. from the distribution of episodic returns (which would invalidate the safety test and the probabilistic safety guarantee).

For generalization gridworld, the optimization algorithm used is an actor-critic with eligibility traces (see Sutton & Barto (2018), Section 13.5), a tabular state-action value function, and a softmax policy. The optimization algorithm's hyperparameters were: actor step size = 0.137731127022912, critic step size = 0.31442900745165847, $\gamma = 1.0$, and $\lambda = 0.23372572419318238$. We also experimented with REINFORCE (Williams, 1992), and the outcomes were nearly identical, with all guarantees holding.

For DAS1, the optimization algorithm used was REINFORCE (Williams, 1992), with eligibility traces, a linear function approximator using the Fourier basis (Konidaris et al., 2011), and a softmax policy. The optimization algorithm's hyperparameters were: $\gamma = 1.0$, step size $= 5.736495301650456(10^{-6})$, $\lambda = 0.9082498629094096$, order = 2, and maximum coupled variables = 2.

In practice, one would tune the hyperparameters of the optimization algorithm using the training set (and *not* the safety set). For the purposes of these experiments, we used the entire underlying distribution μ to tune the hyperparameters of the optimization algorithm. This does not break any of our guarantees, since we have access to the entire true underlying distribution. This methodology is also necessary, since, for each trial, the training set is different, and it is not computationally feasible to do a hyperparameter search for each of the hundreds of thousands of trials represented by our eight experiments.



(a) Hoeffding and t-test DAS1 results

(b) CVaR DAS1 Results

Figure 7. See the caption of the Figure 1 for a general description of the plots. These plots were generated using 100 trials per data point. Where the MDP CVaR curves are not visible, they are overlapping with episodic CVaR curves. Notice that the CVaR HCGAs have lower plotted return than the standard RL algorithm. As discussed in supplementary material Section K.1, this is an artifact of plotting the $J(\text{NSF}) = j$. Alternate plots excluding these trials are given in Figure 11 in the supplementary material. These alternative plots show that, excluding NSF trials, the average returns of CVaR HCGAs are higher than those of the standard RL algorithm.

Again, in practice, when one wishes to apply an HCGA and has access to M_train and M_safety, but not M_test or the true distribution, μ, it is important to do hyperparameter tuning *only* on M_train and not M_safety (otherwise the safety guarantees will be invalid). It is also computationally feasible to do this in practice as this search will only have to be run once (rather than hundreds of thousands of times that would have been required by our experiments).

K.2. DAS1 Results

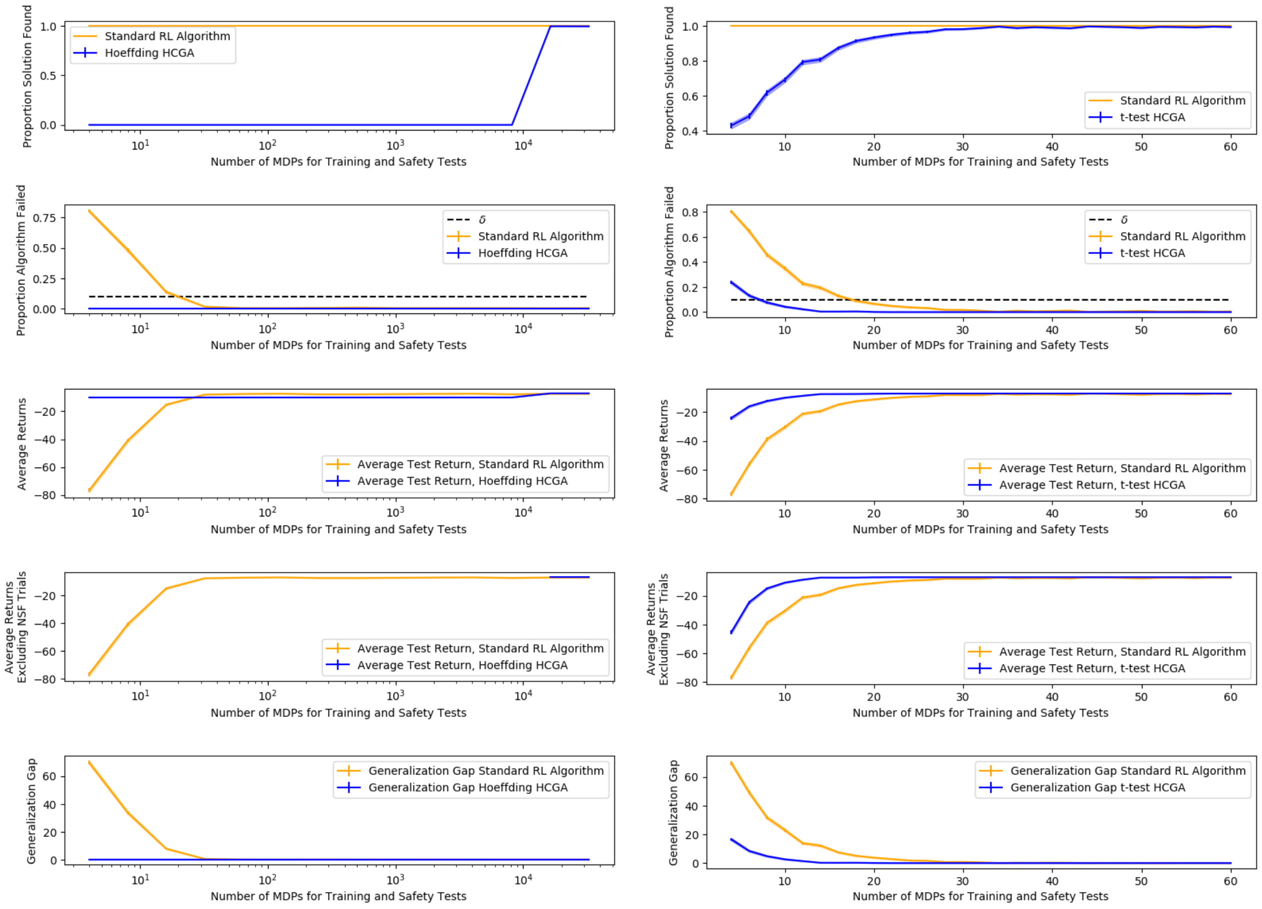
For comparison purposes, the results of the DAS1 experiments in Figure 7 are presented in a layout similar to that of Figure 1 in Section 9. Both environments' results are presented more completely in Section K.3.

K.3. Full Results

In this section, we provide the results for all eight experiments (four HCGAs run for two environments) in eight individual plots. This shows the results of individual experiments more clearly, and allows us to plot the t-test HCGA experiments on a more appropriate linear scale (as opposed to the initial portion of the log scale they are plotted on in Figures 1a and 7a).

We also provide an additional plot for each experiment: the return with NSF trials excluded. That is, instead of plotting the return of NSF trials as j, we exclude those trials from the plot. Notice that, in these alternate plots, some curves do not begin until after $|M_\text{acc}|$ is sufficiently large to cause algorithms to return solutions that are not NSF.

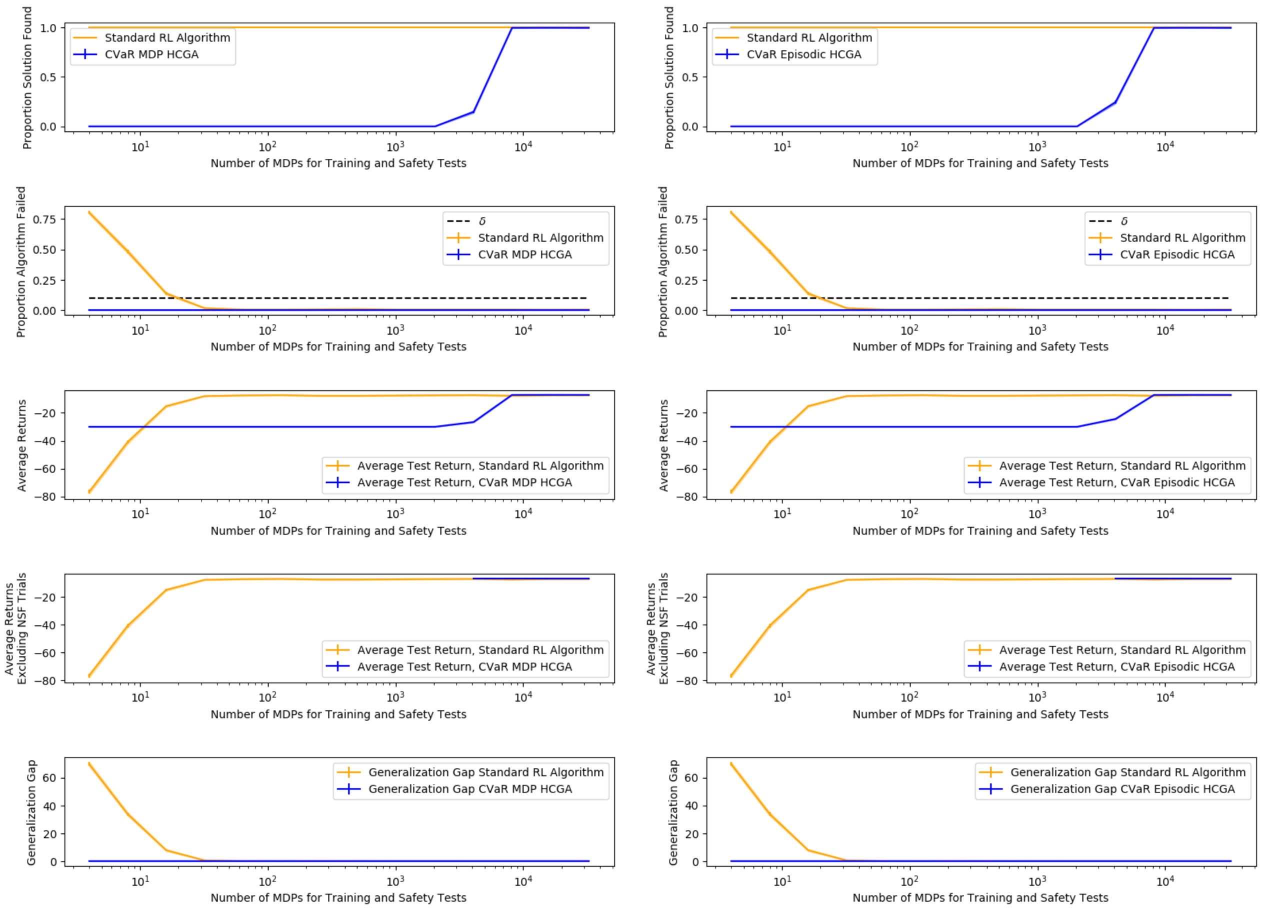
The plots are shown in Figures 8, 9, 10, and 11.



(a) Full Hoeffding Gridworld Results

(b) Full t-test Gridworld Results

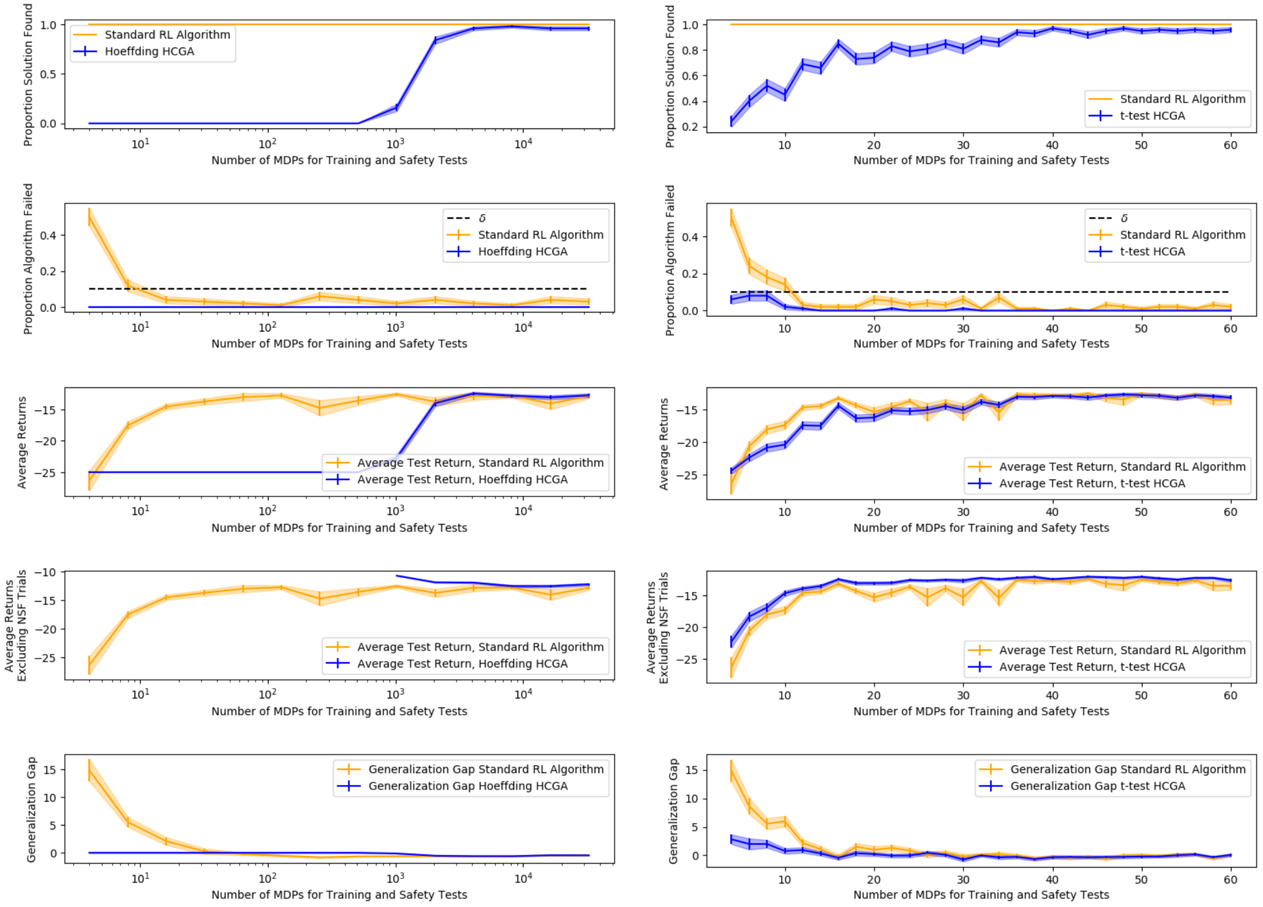
Figure 8. Full Hoeffding and t-test Gridworld Results



(a) Full CVaR MDP HCGA Gridworld Results

(b) Full CVaR Episodic HCGA Gridworld Results

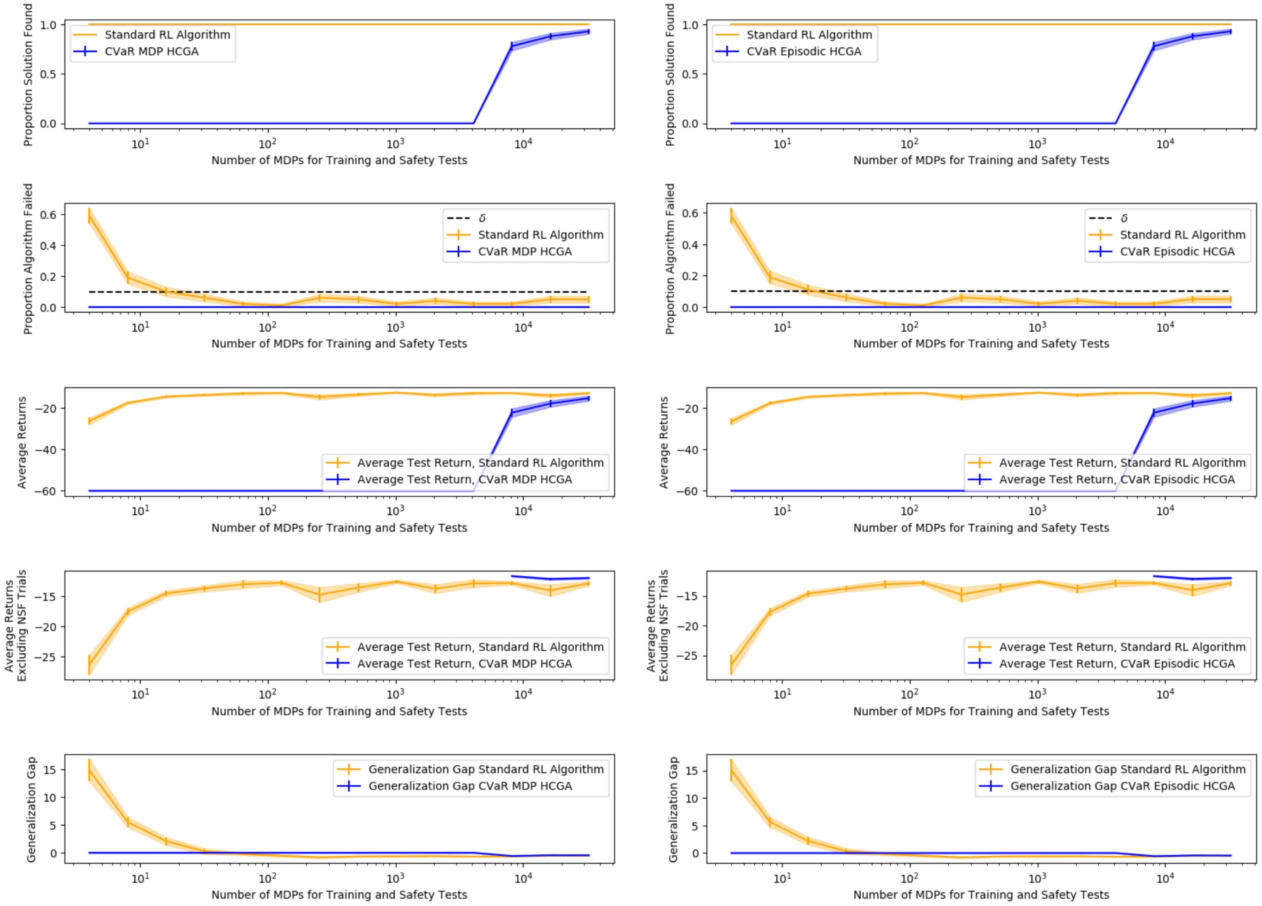
Figure 9. Full CVaR HCGAs Gridworld Results



(a) Full Hoeffding DAS1 Results

(b) Full t-test DAS1 Results

Figure 10. Full Hoeffding and t-test DAS1 Results



(a) Full CVaR MDP HCGA DAS1 Results

(b) Full CVaR Episodic HCGA DAS1 Results

Figure 11. Full CVaR HCGAs DAS1 Results

L. Example HCGA

In this section, we give the algorithm represented by the bounding function defined in (2). This serves as an example of how to apply bounding functions to Algorithm 1 to form a complete HCGA.

Algorithm 3 Expected Return HCGA, Hoeffding Variant

```
Input  : Feasible set $\Theta$, a set of MDPs $M_{\text{acc}}$, user-defined threshold $j$, and probability $1-\delta$.
Output : $\theta \in \Theta \cup \{\text{NSF}\}$
Partition $M_{\text{acc}}$ into two data sets, $M_{\text{train}}$ and $M_{\text{safety}}$;
Compute a $\theta_c \in \text{argmax}_{\theta \in \Theta} J_{M_{\text{train}}}(\theta)$;
if $J_{M_{\text{safety}}}(\theta_c) - \sqrt{\frac{\ln(1/\delta)}{2|M_{\text{safety}}|}} \geq j$ then return $\theta_c$;
else return No Solution Found;
```

M. Expected Return HCGAs with Control Variates: Results and Analysis

In this section, we present and analyze empirical results for the t-test HCGA with control variates. In Sections M.2 and M.3, for procedural gridworld and DAS1 respectively, we demonstrate empirically that the use of control variates with HCGAs reduces the standard deviation of the mean estimates, and that this modification does not violate the HCGAs' safety guarantees. We analyze these results and make a prediction about the kinds of environment distributions for which control variates will significantly reduce the rate at which HCGAs return NSF. Finally, in Section M.4, we use this prediction to construct and study an MDP distribution. For this MDP distribution, control variates result in a significant decrease in the proportion of trials for which NSF is returned.

M.1. Experiment Details

We only study the t-test HCGA in this section; recall the bounds for the two expected value HCGAs above:

$$\text{Hoeffding: } b(\theta, M_{\text{safety}}, \delta) \coloneqq J_{M_{\text{safety}}}(\theta) - \sqrt{\ln(1/\delta)/(2|M_{\text{safety}}|)}.$$

$$\text{t-test: } b(\theta, M_{\text{safety}}, \delta) \coloneqq J_{M_{\text{safety}}}(\theta) - \frac{\hat{\sigma}_J(\theta, M_{\text{safety}})\mathtt{t}_{1-\delta,|M_{\text{safety}}|-1}}{\sqrt{|M_{\text{safety}}|}}.$$

We study only the t-test HCGA because it has a standard deviation term in the bound that is desirable to minimize, and the Hoeffding HCGA does not have such a term. Ignoring computational cost, using control variates for the Hoeffding HCGA could be considered a strict improvement over not using control variates: control variates will reduce the variance of the mean estimates without compromising the safety guarantees. However, control variates will not usually substantially affect the accuracy of the mean estimate and so cannot be expected to improve the Hoeffding HCGA significantly (unlike the t-test HCGA, which, because of the standard deviation term in the bound, may be substantially improved by control variates).

Recall the two methods for choosing a c value: estimate $c = \frac{\mathbf{E}\left[\left(J_{M_i}(\theta) - \mathbf{E}[J_{M_k}(\theta)|M_k \sim \mu]\right)\left(\bar{v}_\theta(p_i) - \mathbf{E}[\bar{v}_\theta(p_j)|M_j \sim \mu]\right)\middle| M_i \sim \mu\right]}{\mathbf{E}\left[\left(\bar{v}_\theta(p_{i'}) - \mathbf{E}[\bar{v}_\theta(p_{j'})|M_{j'} \sim \mu]\right)^2 \middle| M_{i'} \sim \mu\right]}$, or choose $c = 1$. Below, we refer to these variants as the *optimal c estimation* variant, and the $c = 1$ variant, respectively. We study both variants below.

In all experiments in this section, $|M_{\text{safety}}| \geq 32$ (so the horizontal axis, $|M_{\text{acc}}|$, begins at $2|M_{\text{safety}}| = |M_{\text{acc}}| = 64$). This is because the t-test bound assumes that the performances of θ_c for the MDPs in M_{safety} are normally distributed. This assumption may not be reasonable, particularly for small values of $|M_{\text{safety}}|$. However, by the central limit theorem, it is often a reasonable assumption for large values of $|M_{\text{safety}}|$.

For simplicity, we use the k-nearest neighbors algorithm for the control variate. We chose $k = 3$ based on intuition, and did not tune or try any other values of this hyperparameter (note that, regardless of the value of this hyperparameter, the safety guarantees will hold and the Z_i values based on the control variate will be unbiased estimators of $J_\mu(\theta)$). As mentioned above, the control variate function approximator, supervised learning algorithm, optimizer, and hyperparameters can be arbitrary. For example, a deep neural network or linear function approximator trained with stochastic gradient descent may also be suitable for many problem settings.

M.2. Control Variates: Generalization Gridworld Results

Consider Figure 12, which shows the generalization gridworld results for HCGAs using control variates. Notice that both variants reduce the standard deviation of the mean estimators compared to the HCGA with no control variate (first plot). However, this fact does not help the HCGAs return more solutions for this environment (third plot), since

$$J_{M_{\text{safety}}}(\theta) \gg \frac{\hat{\sigma}_J(\theta, M_{\text{safety}})\mathtt{t}_{1-\delta,|M_{\text{safety}}|-1}}{\sqrt{|M_{\text{safety}}|}}.$$

In other words, the $\frac{\hat{\sigma}_J(\theta, M_{\text{safety}})\mathtt{t}_{1-\delta,|M_{\text{safety}}|-1}}{\sqrt{|M_{\text{safety}}|}}$ term is already insignificant compared to $J_{M_{\text{safety}}}(\theta)$, so decreasing it more using control variates does not help significantly in practice. Also notice that the optimal c estimates are approximately one (second plot); this fact explains the similar performance of the two control variate variants, and adds empirical support to the theory that $c = 1$ (Corollary 1) is a useful rule. Finally, notice that the control variate HCGAs do not violate the safety guarantees (fourth plot).

As shown in the fourth plot of Figure 12, lowering the standard deviation of the mean estimators for the safety test does not significantly reduce the rate at which the HCGA returns NSF for the generalization gridworld. This is because, for all environments in the distribution, $J(\theta^*)$ is the same value, where θ^* are the parameters of the optimal policy. That is, for all

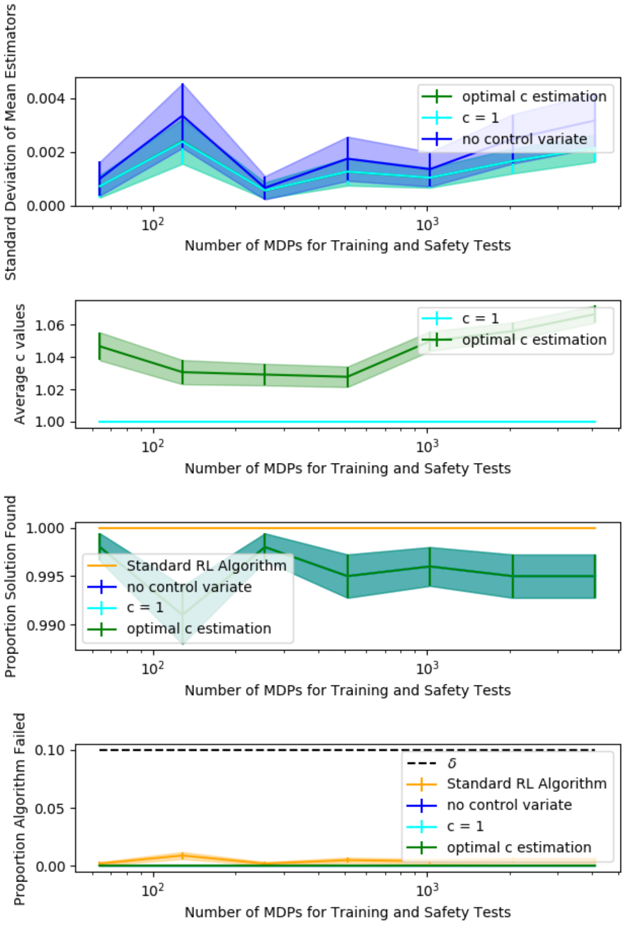


Figure 12. Results for control variates for generalization gridworld. Each location on the horizontal axis corresponds to 1000 trials. Where the $c = 1$ control variate curve is not visible, it is overlapping with the optimal c estimation curve.

MDPs $m \in \operatorname{supp}(\mu)$, $J_m(\theta^*) = -7$ (before return/objective normalization). In this environment, when the HCGA has enough data to pass the safety test, the HCGA tends to learn a policy very close to the optimal policy. Because the standard deviation of the objectives, $\hat{\sigma}_J$, is close to zero, the reduction of the standard deviation of the mean estimates is not helpful in practice for this environment. As we show below, control variates may help more in practice for environment distributions with more variation in the objective functions of their MDPs (given typical candidate policies).

Note, for the generalization gridworld results only: for calculated c values in approximately 0.1% of trials, the denominator of the c calculation is exactly 0, resulting in an undefined value of c. In these rare cases, we set $c = 1$ based on Corollary 1. Notice that, for these rare trials, this method disregards the step in Algorithm 2 that says "ensure that $\bar{v}_{\theta_c}$ is not a constant"; the control variate is a constant in these cases, which results in the calculated c value being undefined. Because our goal is to show properties of these algorithms across many trials (not to apply them to a real-world problem in practice), we use this method for this subsection only. Since this denominator value is never zero in the environments below, this method is only necessary for this generalization gridworld subsection, not for the subsections below. (This is because, given typical candidate policies, the standard deviation between the objectives of different MDPs tends to be at least an order of magnitude lower for generalization gridworld than for the other two environments below. This low standard deviation can sometimes result in a constant control variate, since the function the control variate is supposed to approximate is a constant or nearly a constant.)

M.3. Control Variates: DAS1 Results

Figure 13 shows the DAS1 results for HCGAs using control variates; the results are similar to the generalization gridworld results. Both variants reduce the variance of the mean estimators compared to the HCGA with no control variate. Also notice that, once again, the optimal c estimates are approximately one (second plot), which adds further empirical support to the theory that $c = 1$ (Corollary 1) is a useful rule. Additionally, notice that the control variate HCGAs do not violate the safety guarantees (fourth plot).

Like the results for the generalization gridworld, these results do not show that the control variates result in a significant increase in the proportion of trials in which a solution is found.

Consideration of these results naturally leads to the observation that control variates might be more useful in practice for an environment distribution with much higher standard deviation between the objectives of different MDPs (for optimal or near optimal policies). We explore this observation in the next subsection.

M.4. Control Variates: Stochastic Generalization Gridworld Results

We modified the generalization gridworld MDP distribution so that optimal or near optimal policies would result in higher standard deviation between the objectives of different MDPs: Half of MDPs in the modified distribution are exactly like the MDPs in the unmodified distribution. The other half of MDPs have a stochasticity in their transition functions of 0.5. This means that with probability 0.5, the agent would move as in the normal generalization gridworld transition function. With probability 0.5, the environment will ignore the agent's action, and force it to move in a random direction (or attempt to move in that direction, since the boundaries of the environments or the cliffs may interfere). There are four directions, so the result is that the agent has a 0.625 probability of moving in its "intended" direction, and a 0.125 probability of moving in one of the other three directions. We call this the *stochastic generalization gridworld*.

To account for the changed dynamics, we changed the definition of safety for this environment by decreasing the value of a safe j and increasing the probability with which a safe solution must be returned: $1 - \delta := 0.99$ and $j := -23$ (before return/objective normalization).

As shown in Figure 14, the results match our hypothesis; control variates are more useful in practice for an MDP distribution like the stochastic generalization gridworld. The "proportion solution found" plot shows that, for $|M_{\text{acc}}| = 64$ and $|M_{\text{acc}}| = 128$, the control variates substantially reduce the probability that NSF is returned. As in the experiments above, the safety guarantees were not violated, and the optimal c estimation value was approximately 1.

These results empirically confirm our theory that control variates can reduce the value of the standard deviation of the mean estimates for expected return HCGAs, and that this modification does not violate the HCGA safety guarantees. Furthermore, these results show that the control variate extension may be particularly useful for environments which have high variance in objectives between MDPs (for a typical candidate policy).

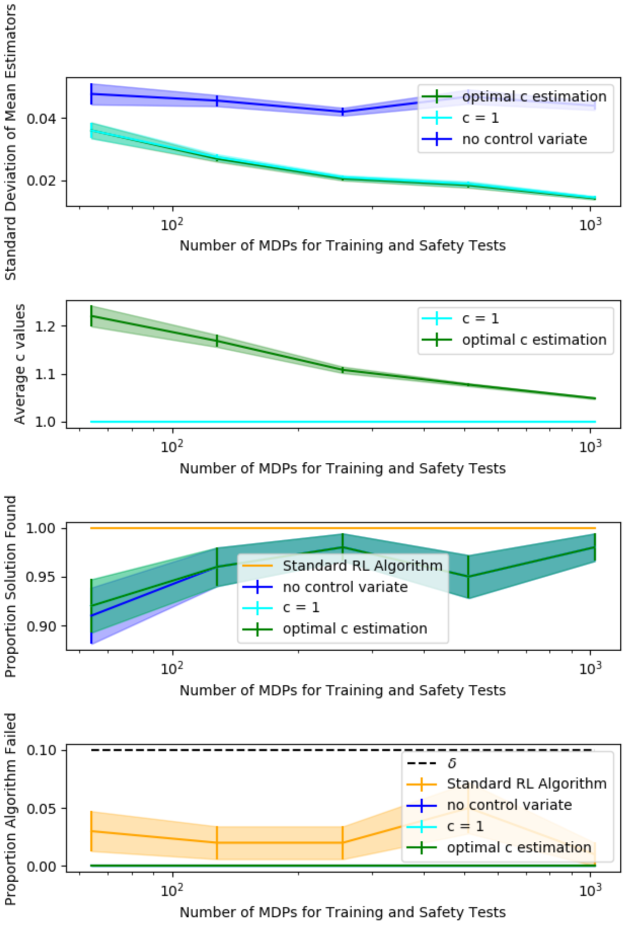


Figure 13. Results for control variates for DAS1. Each location on the horizontal axis corresponds to 100 trials. Where the $c = 1$ control variate curve is not visible, it is overlapping with the optimal c estimation curve.

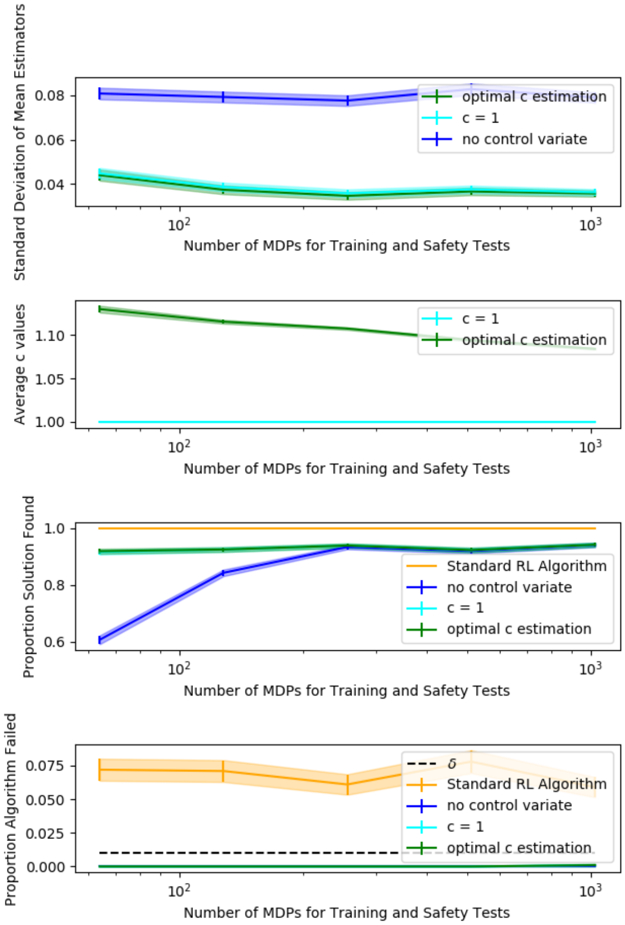


Figure 14. Results for control variates for stochastic generalization gridworld. Each location on the horizontal axis corresponds to 1000 trials. Where the $c = 1$ control variate curve is not visible, it is overlapping with the optimal c estimation curve.