A Appendix

A.1 Proof of Theorem 1

We begin with an introduction to the PAC-Bayes framework, and then provide a complete proof of Theorem 1. PAC-Bayes provides an upper bound on the expected cost of deploying a policy distribution P on environments E drawn from an unknown distribution $\mathcal{D}$, i.e., $\mathbb{E}_{E\sim\mathcal{D}}\mathbb{E}_{\pi\sim P}C_E(\pi)$. This upper bound only depends on the cost of deploying P in a finite set of training environments $S \sim \mathcal{D}^m$, i.e., the training cost $\mathbb{E}_{\pi\sim P}C_{E_i}(\pi)$, and a regularizer which depends on the KL-divergence between P and a prior P_0 that is chosen before observing S; note that P_0 need not be a Bayesian prior. The following is the PAC-Bayes bound that was presented in [26] and tightened in [49]:

Theorem 4 (PAC-Bayes Bound [26]) *For any distribution over environments $\mathcal{D}$, data-independent prior distribution P_0, cost C bounded in $[0,1]$, $m \geq 8$, and $\delta \in (0,1)$, with probability at least $1-\delta$ over a sampling of $S \sim \mathcal{D}^m$, the following holds for all posterior distributions P:*

$$\mathop{\mathbb{E}}_{E\sim\mathcal{D}}\ \mathop{\mathbb{E}}_{\pi\sim P} C_E(\pi) \leq \sum_{i=1}^{n} \mathop{\mathbb{E}}_{\pi\sim P} C_{E_i}(\pi) + \sqrt{\frac{D_{\mathrm{KL}}(P\|P_0) + \ln\frac{2\sqrt{m}}{\delta}}{2m}} \tag{6}$$

where D_{KL} is the KL-divergence.

The above theorem and the forthcoming PAC-Bayes theorems in this section are presented for policy learning instead of supervised learning using the reduction provided in [40]. Note that this bound provides a guarantee for a *distribution* over policies rather than a specific policy. This allows for a regularizer dependent on the KL-divergence between the prior and posterior distributions rather than one which is a direct expression of the complexity of the policy space (such as the VC-dimension). However, this creates a challenge for calculating the upper bound, which requires computing an expectation over $\pi \sim P$, or using potentially-loosening sample convergence bounds. Thus, we make use of the recent work which provides a framework for derandomized PAC-Bayes bounds (i.e. bounds which hold for a sampling of policy π rather than an expectation over $\pi \sim P$) [44]. The following is a general theorem for formulating the derandomized PAC-Bayes bounds:

Theorem 5 (Pointwise PAC-Bayes Bound [44]) *For any positive function ϕ, distribution $\mathcal{D}$, prior distribution P_0, and $\delta \in (0,1)$, with probability $1-\delta$ over a sampling of $S \sim \mathcal{D}^m$ and $\pi \sim P$, the following holds for any posterior distribution P:*

$$\frac{\alpha}{\alpha-1}\ln(\phi(\pi,S)) \leq D_\alpha(P\|P_0) + \ln\left(\frac{1}{(\delta/2)^{\frac{\alpha}{\alpha-1}+1}} \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\ \mathop{\mathbb{E}}_{\pi'\sim P_0} \phi(\pi',S')^{\frac{\alpha}{\alpha-1}}\right) \tag{7}$$

where P is the output of algorithm A on the training data S, i.e. $P := A(P_0, S)$ and D_α is the Rényi divergence.

Now we can proceed with the statement and proof.

Theorem 1 *For any distribution $\mathcal{D}$, prior distribution P_0, $\delta \in (0,1)$, cost bounded in $[0,1]$, and deterministic algorithm A which outputs the posterior distribution P we have the following:*

$$\mathop{\mathbb{P}}_{(S,\pi)\sim(\mathcal{D}^m\times P)}\left[C_{\mathcal{D}}(\pi) \leq C_S(\pi) + \sqrt{\frac{D_2(P\|P_0) + \ln\frac{2\sqrt{m}}{(\delta/2)^3}}{2m}}\right] \leq 1-\delta \tag{8}$$

where D_2 is the Rényi Divergence for $\alpha = 2$.

Proof. We begin with the statement in Theorem 5, which is proved in [44]. Let $\alpha = 2$ and $\phi(\pi,S) = \exp[\frac{\alpha-1}{\alpha}mD_{\mathrm{KL}}(C_S(\pi)\|C_{\mathcal{D}}(\pi))]$. Thus, we have the following with at least probability $1-\delta$ over the random choice $S \sim \mathcal{D}^m$ and $\pi \sim P$:

$$D_{\mathrm{KL}}(C_S(\pi)\|C_{\mathcal{D}}(\pi)) \leq \frac{1}{m}\left[D_2(P\|P_0) + \ln\left(\frac{1}{(\delta/2)^3} \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^m}\ \mathop{\mathbb{E}}_{\pi'\sim P_0} e^{mD_{\mathrm{KL}}(C_{S'}(\pi')\|C_{\mathcal{D}}(\pi'))}\right)\right] \tag{9}$$

From [49], we can upper bound $\mathbb{E}_{S'\sim\mathcal{D}^m}\ \mathbb{E}_{\pi'\sim P_0}\ e^{mD_{\mathrm{KL}}(C_{S'}(\pi')\|C_{\mathcal{D}}(\pi'))}$ by $2\sqrt{m}$ when $m \geq 8$. This gives us the following bound

$$D_{\mathrm{KL}}(C_S(\pi)\|C_{\mathcal{D}}(\pi)) \leq \frac{1}{m}\left[D_2(P\|P_0) + \ln\frac{2\sqrt{m}}{(\delta/2)^3}\right]. \tag{10}$$

We then apply the Pinkser's inequality, i.e. $D_{\mathrm{KL}}(p\|q) \leq c \implies q \leq p + \sqrt{c/2}$, which results in Inequality (2). Note that we could also use a quadratic version of the upper bound for the KL divergence between two distributions and produce an upper bound analogous to the one presented in [37]. □

A.2 Proof of Theorem 2

For the readers' convenience, we restate Theorem 2 here and provide a detailed proof.

Theorem 2 *Let $\mathcal{D}$ be the training distribution and P be the posterior distribution on the space of policies obtained through the training procedure described in Sec. 4.1. Let $S' \sim \mathcal{D}'^n$ be a test dataset, $p(S')$ be the p-value for S' defined in Definition 1, and $\delta \in (0,1)$. Then,*

$$\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2)] \geq 1-\delta \ , \tag{11}$$

where $\overline{\tau}(S) := \max\{C_{S'}(\pi) - \overline{C}_\delta(\pi, S), 0\}$.

Proof. We prove this theorem by considering two cases: when the PAC-Bayes cost inequality holds, i.e., $C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)$, and when it does not, i.e., $C_{\mathcal{D}}(\pi) > \overline{C}_\delta(\pi, S)$; the two cases are considered in (12)-(14). In the latter case, we cannot say anything about the p-value, while in the former case, which holds with probability at least $1-\delta$, we show in (15)-(26) that $p(S') \leq \exp(-2n\overline{\tau}(S)^2)$.

Let us begin the proof by conditioning $\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2)]$ as follows:

$$\begin{aligned}
&\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2)] \qquad (12)\\
&= \underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \mid C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)] \underbrace{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)]}_{\geq 1-\delta \text{ (from Theorem 1)}}\\
&+ \underbrace{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \mid C_{\mathcal{D}}(\pi) > \overline{C}_\delta(\pi, S)]}_{\geq 0} \underbrace{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi) > \overline{C}_\delta(\pi, S)]}_{\geq 0} \qquad (13)\\
&\geq \underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \mid C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)](1-\delta) \qquad (14)
\end{aligned}$$

Now, we claim:

$$\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \mid C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)] = 1 \ , \tag{15}$$

which on using in (14) completes the proof of this theorem. The remainder of this proof is dedicated to establishing the claim in (15).

We are given

$$C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S) \ . \tag{16}$$

From Definition 1, we have

$$\begin{aligned}
p(S') &= \underset{\hat{S}\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\hat{S}}(\pi) \geq C_{S'}(\pi) \mid C_{\mathcal{D}'}(\pi) \leq C_{\mathcal{D}}(\pi)] && (17)\\
&= \underset{\hat{S}\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\hat{S}}(\pi) - \overline{C}_\delta(\pi, S) \geq C_{S'}(\pi) - \overline{C}_\delta(\pi, S) \mid C_{\mathcal{D}'}(\pi) \leq C_{\mathcal{D}}(\pi)] && (18)\\
&= \underset{\hat{S}\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\hat{S}}(\pi) - \overline{C}_\delta(\pi, S) \geq \overline{\tau} \mid C_{\mathcal{D}'}(\pi) \leq C_{\mathcal{D}}(\pi)] \ . && (19)
\end{aligned}$$

From (16) and the assumption that the null hypothesis holds in (19), it follows that $C_{\mathcal{D}'}(\pi) \leq \overline{C}_\delta(\pi, S)$, which ensures that the following implication holds for $\overline{\tau}$ defined in the statement of the theorem:

$$C_{\hat{S}}(\pi) - \overline{C}_\delta(\pi, S) \geq \overline{\tau} \implies C_{\hat{S}}(\pi) - C_{\mathcal{D}'}(\pi) \geq \overline{\tau} \ . \tag{20}$$

Therefore, if $\overline{\tau} > 0$ we have that

$$\underset{\hat{S}\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\hat{S}}(\pi) - \overline{C}_\delta(\pi, S) \geq \overline{\tau} \mid C_{\mathcal{D}'}(\pi) \leq C_{\mathcal{D}}(\pi)] \leq \underset{\hat{S}\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\hat{S}}(\pi) - C_{\mathcal{D}'}(\pi) \geq \overline{\tau}] \leq \exp(-2n\overline{\tau}^2) \ ,$$

where the last upper bound follows from Hoeffding's inequality. Hence, for $\overline{\tau} > 0$, using the above in (19) gives

$$p(S') \leq \exp(-2n\overline{\tau}^2) \ . \tag{21}$$

If $\overline{\tau} = 0$, the vacuous bound holds:

$$p(S') \leq 1 = \exp(-2n0) = \exp(-2n\overline{\tau}^2) \ . \tag{22}$$

Combining the two cases for $\overline{\tau} > 0$ in (21) and $\overline{\tau} = 0$ in (22) gives us the following implication:

$$C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S) \implies p(S') \leq \exp(-2n\overline{\tau}(S)^2) \ . \tag{23}$$

Now, we expand the left-hand side of (15) using the definition of conditional probability:

$$\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \mid C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)] \tag{24}$$

$$= \frac{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \ \wedge \ C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)]}{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)]} \tag{25}$$

From (23), we know that $\{(S,\pi) \mid C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)\} \subseteq \{(S,\pi) \mid p(S') \leq \exp(-2n\overline{\tau}(S)^2)\}$, therefore, $\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \ \wedge \ C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)] = \underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)]$ which on using in (25) gives the following:

$$\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[p(S') \leq \exp(-2n\overline{\tau}(S)^2) \mid C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)] = \frac{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)]}{\underset{(S,\pi)\sim(\mathcal{D}^m \times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi) \leq \overline{C}_\delta(\pi, S)]} = 1, \tag{26}$$

completing the proof of the claim (15) as well as the theorem. □

A.3 Proof of Theorem 3

For the readers' convenience, we restate Theorem 3 here and provide a detailed proof.

Theorem 3 *Let $\mathcal{D}$ be the training distribution, $\mathcal{D}'$ be the test distribution, and P be the posterior distribution on the space of policies obtained through the training procedure described in Sec. 4.1. Let $\delta, \delta' \in (0,1)$ such that $\delta + \delta' < 1$, $\gamma := \sqrt{\frac{\ln(1/\delta')}{2n}}$, and $\Delta C := C_{S'}(\pi) - \gamma - \overline{C}_\delta(\pi, S)$. Then,*

$$\underset{(S,\pi,S')\sim(\mathcal{D}^m \times P \times \mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}'}(\pi) - C_{\mathcal{D}}(\pi) \geq \Delta C] \geq 1 - \delta - \delta' \ . \tag{27}$$

Proof. To lower bound the difference between $C_{\mathcal{D}'}(\pi)$ and $C_{\mathcal{D}}(\pi)$ with high probability we obtain a lower bound on $C_{\mathcal{D}'}(\pi)$ which holds with probability at least $1 - \delta'$ using Hoeffding's inequality in (28)-(31). Then we use this bound with the PAC-Bayes bound (2) which holds with probability at least $1 - \delta$ to obtain (27) by following the steps in (32)-(37).

Let γ be defined as in the statement of the theorem, then, using the independence of $\mathcal{D}'^n$ from $\mathcal{D}^m \times P$, we can write[2]

$$\underset{(S,\pi,S')\sim(\mathcal{D}^m \times P \times \mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}'}(\pi) \geq C_{S'}(\pi) - \gamma]$$

$$= \int_{(S,\pi)} \underset{S'\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\mathcal{D}'}(\pi) \geq C_{S'}(\pi) - \gamma \mid S, \pi] d(\mathcal{D}^m \times P)(S, \pi) \ . \tag{28}$$

For *any* given (S, π), we can apply Hoeffding's inequality to get:

$$\underset{S'\sim\mathcal{D}'^n}{\mathbb{P}}[C_{\mathcal{D}'}(\pi) \geq C_{S'}(\pi) - \gamma] = \underset{S'\sim\mathcal{D}'^n}{\mathbb{P}}[C_{S'}(\pi) - C_{\mathcal{D}'}(\pi) \leq \gamma] \geq 1 - \exp(-2n\gamma^2) = 1 - \delta' \ . \tag{29}$$

---

[2] Note that $C_{\mathcal{D}'}(\pi)$ and $C_{S'}(\pi)$ implicitly depend on S because the posterior distribution P, from which π is sampled, is trained on S.

Using (29) in (28) we get that:

$$\underset{(S,\pi,S')\sim(\mathcal{D}^m\times P\times\mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}'}(\pi)\geq C_{S'}(\pi)-\gamma]\geq\int_{(S,\pi)}(1-\delta')d(\mathcal{D}^m\times P)(S,\pi) \tag{30}$$

$$=(1-\delta')\int_{(S,\pi)}d(\mathcal{D}^m\times P)(S,\pi)=1-\delta' \ . \tag{31}$$

Now, observe that

$$C_{\mathcal{D}}(\pi)\leq\overline{C}_{\delta}(\pi,S)\ \wedge\ C_{\mathcal{D}'}(\pi)\geq C_{S'}(\pi)-\gamma\implies C_{\mathcal{D}'}(\pi)-C_{\mathcal{D}}(\pi)\geq C_{S'}(\pi)-\gamma-\overline{C}_{\delta}(\pi,S) \ . \tag{32}$$

From the implication (32), it follows that

$$\underset{(S,\pi,S')\sim(\mathcal{D}^m\times P\times\mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}'}(\pi)-C_{\mathcal{D}}(\pi)\geq C_{S'}(\pi)-\gamma-\overline{C}_{\delta}(\pi,S)] \tag{33}$$

$$\geq\underset{(S,\pi,S')\sim(\mathcal{D}^m\times P\times\mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}}(\pi)\leq\overline{C}_{\delta}(\pi,S)\ \wedge\ C_{\mathcal{D}'}(\pi)\geq C_{S'}(\pi)-\gamma] \tag{34}$$

Now using the Fréchet inequality $\mathbb{P}[E_1\wedge E_2]\geq\mathbb{P}[E_1]+\mathbb{P}[E_2]-1$ (where E_1 and E_2 are arbitrary random events) on (34) we obtain:

$$\underset{(S,\pi,S')\sim(\mathcal{D}^m\times P\times\mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}}(\pi)\leq\overline{C}_{\delta}(\pi,S)\ \wedge\ C_{\mathcal{D}'}(\pi)\geq C_{S'}(\pi)-\gamma] \tag{35}$$

$$\geq\underset{(S,\pi)\sim(\mathcal{D}^m\times P)}{\mathbb{P}}[C_{\mathcal{D}}(\pi)\leq\overline{C}_{\delta}(\pi,S)]+\underset{(S,\pi,S')\sim(\mathcal{D}^m\times P\times\mathcal{D}'^n)}{\mathbb{P}}[C_{\mathcal{D}'}(\pi)\geq C_{S'}(\pi)-\gamma]-1 \tag{36}$$

$$\geq1-\delta-\delta' \ , \tag{37}$$

where the last inequality follows by using (2) and (31) in (36). Finally, using (37) in (34) completes the proof. □

A.4 Training with Backpropogation

In this section, we describe a method to minimize the upper bound in Theorem 1 using backpropogation. We make use of multivariate Gaussian distributions $\mathcal{N}_{\psi}$ with diagonal covariance $\Sigma_s:=\text{diag}(s)$ where $\psi:=(\mu,\log s)$. When training the posterior distribution P, we would like to take gradient steps directly with respect to ψ. However, this would require backpropagation through $\mathbb{E}_{w\sim\mathcal{N}_{\psi}}C_E(\pi_w)$. We follow a similar procedure as in [28] and achieve the desired result of minimizing the upper bound in Inequality (2) using an unbiased estimate of $\mathbb{E}_{w\sim\mathcal{N}_{\psi}}C_E(\pi_w)$:

$$\frac{1}{k}\sum_{i=1}^{k}C_E(\pi_{w_i}),\quad w_i\sim\mathcal{N}_{\psi}\ \forall\ i\in\{1,2,\ldots,k\} \ . \tag{38}$$

The resulting approach is presented in Algorithm 1. Note that the algorithm must be deterministic in order to maintain the assumptions of Theorem 1. We achieve this by training with a fixed seed for generating random numbers. Additionally, note that the backpropagation requires a gradient taken through $D_2(\mathcal{N}_{\psi}\|\mathcal{N}_{\psi_0})$. We make use of the analytical form for the Rényi divergence between two multivariate Gaussian distributions, presented in [50], in order to tractably compute the gradients.

$$D_2(\mathcal{N}_{\psi}\|\mathcal{N}_{\psi_0})=D_2\big(\mathcal{N}(\mu,\Sigma_s)\|\mathcal{N}(\mu_0,\Sigma_{s_0})\big)=(\mu-\mu_0)^T\Sigma_2(\mu-\mu_0)-\frac{1}{2}\ln\frac{|\Sigma_2||\Sigma_s|}{|\Sigma_{s_0}|^2} \ , \tag{39}$$

where $\Sigma_2=2\Sigma_{s_0}-\Sigma_s$. We also note that there is a restriction on how far the posterior's variance can drift from the prior. The following expression must be satisfied for $D_2(\mathcal{N}_{\psi}\|\mathcal{N}_{\psi_0})$ to be finite [50]:

$$2\Sigma_s^{-1}-\Sigma_{s_0}^{-1}\succ0. \tag{40}$$

In practice, we project any problematic variances into the range of allowable variances.

After training, since we have used the pointwise PAC-Bayes bound in Theorem 1, we compute the upper bound with a single $w\sim\mathcal{N}_{\psi}$ in contrast to traditional PAC-Bayes bounds. Thus, the resulting policy π_w is deterministic and applicable in a broad range of settings, including ones which require a pre-trained network. The resulting policy carries a PAC-Bayes guarantee.

Algorithm 1 PAC-Bayes Bound Minimization via Backpropagation

```
**Input**: Fixed prior distribution $\mathcal{N}_{\psi_0}$ over policies, fixed seed for random number generation
**Input**: Training dataset $S$, learning rate $\gamma$
**Output**: Optimized $\psi^*$
while not converged do
    Sample $w_i \sim \mathcal{N}_\psi \; \forall \, i \in \{1, 2, ..., k\}$
    $B \leftarrow \frac{1}{mk} \sum_{E \in S} \sum_{i=1}^{k} C_E(\pi_{w_i}) + \sqrt{R}$
    $\psi \leftarrow \psi - \gamma \nabla_\psi B$
end while
```

A.5 Training with Evolutionary Strategies

To train robot control policies in settings where backpropagation is not feasible (e.g. presence of a "blackbox" in the form of a simulator or robot hardware in the forward pass), we use Evolutionary Strategies (ES) which is a class of blackbox optimizers [45]. ES addresses this challenge by estimating the gradient via a Monte-Carlo estimator:

$$\nabla_\psi C_S(\mathcal{N}_\psi) := \frac{1}{m} \sum_{E \in S} \nabla_\psi \mathop{\mathbb{E}}_{w \sim \mathcal{N}_\psi} [C_E(\pi_w)] = \frac{1}{m} \sum_{E \in S} \mathop{\mathbb{E}}_{w \sim \mathcal{N}_\psi} [C_E(\pi_w) \nabla_\psi \ln \mathcal{N}_\psi(w)] \; . \tag{41}$$

Although we can compute the gradient of the regularizer analytically (as mentioned in App. A.4), using different methods to estimate the gradient of the empirical cost (ES) and the gradient of the regularizer (analytically) results in poor convergence. To alleviate this, we estimate the regularizer's gradient using ES as well by leveraging the expectation form of Rényi divergence in Theorem 1. This takes the following form:

$$\begin{aligned} &\nabla_\psi (C_S(\mathcal{N}_\psi) + \sqrt{R}) \\ &= \frac{1}{m} \sum_{E \in S} \mathop{\mathbb{E}}_{w \sim \mathcal{N}_\psi} \Bigg[\underbrace{\left(C_E(\pi_w) + \frac{\mathrm{e}^{\ln\left(\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)}\right) - D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0})}}{4m\sqrt{R}} \right) \nabla_\psi \ln \mathcal{N}_\psi(w)}_{\tilde{C}_E(w)} \Bigg] \; . \end{aligned} \tag{42}$$

A.5.1 Derivation of (42)

To derive (42), note that

$$\nabla_\psi (C_S(\mathcal{N}_\psi) + \sqrt{R}) = \nabla_\psi C_S(\mathcal{N}_\psi) + \nabla_\psi \sqrt{R} = \nabla_\psi C_S(\mathcal{N}_\psi) + \frac{1}{2\sqrt{R}} \nabla_\psi R \; . \tag{43}$$

From (41) we know the gradient for $\nabla_\psi C_S(\mathcal{N}_\psi)$. In the rest of this derivation, therefore, we will focus on the computing the gradient of the second term.

Note that the Rényi divergence for multivariate Gaussian distributions can be written as:

$$D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0}) = \ln \left(\mathop{\mathbb{E}}_{w \sim \mathcal{N}_{\psi_0}} \left[\left(\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)} \right)^2 \right] \right). \tag{44}$$

Let

$$\eta := \mathop{\mathbb{E}}_{w \sim \mathcal{N}_{\psi_0}} \left[\left(\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)} \right)^2 \right] \; , \tag{45}$$

then, using (44) we have that

$$D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0}) = \ln \eta \tag{46}$$

which allows us to express R as

$$R = \frac{\ln \eta + \ln(2\sqrt{m}/(\delta/2)^3)}{2m} \; . \tag{47}$$

Hence,

$$\nabla_\psi R = \frac{1}{2m\eta}\nabla_\psi \eta = \frac{\mathrm{e}^{-D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0})}}{2m}\nabla_\psi \eta \ , \tag{48}$$

where the last equality follows from (46).

For computing the gradient using ES, we require the cost to be an expectation over the posterior, however, η is an expectation on the prior. To address this we perform a change of measure which gives us the following:

$$\eta = \mathop{\mathbb{E}}_{w\sim\mathcal{N}_{\psi_0}}\left[\left(\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)}\right)^2\right] = \mathop{\mathbb{E}}_{w\sim\mathcal{N}_\psi}\left[\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)}\right] \ . \tag{49}$$

Using (49) in (48) gives us

$$\nabla_\psi R = \frac{\mathrm{e}^{-D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0})}}{2m}\nabla_\psi \mathop{\mathbb{E}}_{w\sim\mathcal{N}_\psi}\left[\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)}\right] = \frac{\mathrm{e}^{-D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0})}}{2m}\mathop{\mathbb{E}}_{w\sim\mathcal{N}_\psi}\left[\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)}\nabla_\psi \ln \mathcal{N}_\psi(w)\right] \ . \tag{50}$$

Using (50) and (41) in (43) and combining the expectation terms gives

$$\nabla_\psi(C_S(\mathcal{N}_\psi) + \sqrt{R}) = \tag{51}$$

$$\frac{1}{m}\sum_{E\in S}\mathop{\mathbb{E}}_{w\sim\mathcal{N}_\psi}\left[\left(C_E(\pi_w) + \frac{\mathrm{e}^{-D_2(\mathcal{N}_\psi || \mathcal{N}_{\psi_0})}}{4m\sqrt{R}}\frac{\mathcal{N}_\psi(w)}{\mathcal{N}_{\psi_0}(w)}\right)\nabla_\psi \ln \mathcal{N}_\psi(w)\right] \ . \tag{52}$$

Finally, we note that the dimensionality d of w can be large, in which case the term $\mathcal{N}_\psi(w)/\mathcal{N}_{\psi_0}(w)$ is numerically unstable because it involves the product of d terms. Hence, we express $\mathcal{N}_\psi(w)/\mathcal{N}_{\psi_0}(w)$ as $\mathrm{e}^{\ln(\mathcal{N}_\psi(w)/\mathcal{N}_{\psi_0}(w))}$ which gives us (42) as the final form of the gradient.

A.5.2 Training algorithm

The gradient of the PAC-Bayes upper bound is estimated from (42). Since Theorem 1 requires the training algorithm to be deterministic, we train with a fixed seed. The psuedo-code for our training is provided in Algorithm 2. After training, a single w is drawn from $\mathcal{N}_{\psi^*}$, which corresponds to a policy π_w, and the derandomized PAC-Bayes bound is computed for this policy.

Algorithm 2 PAC-Bayes Bound Minimization via ES

```
Input: Fixed prior distribution $\mathcal{N}_{\psi_0}$ over policies, fixed seed for random number generation
Input: Training dataset $S$, learning rate $\gamma$
Output: Optimized $\psi^*$
while not converged do
    Sample $w_i \sim \mathcal{N}_\psi \ \forall \ i \in \{1, 2, ..., k\}$
    grad ← $\frac{1}{mk}\sum_{E\in S}\sum_{i=1}^{k}\tilde{C}_E(w_i)$
    $\psi \leftarrow \psi - \gamma \cdot$ grad
end while
```

A.6 Additional Experimental Details and Results

A.6.1 Robotic grasping

Training platform. Training was performed on a `Lambda Blade` server with `2x Intel Xeon Gold 5220R` (96 CPU threads) and 768 GB RAM.

Distributions on the initial position of the mugs. For all datasets, mugs are placed upright on the table with random yaw orientations sampled from the uniform distribution $\mathcal{U}([-\pi \text{ rad}, \pi \text{ rad}])$. The following distributions on the mug's placement were used to generate the plot in Fig. 2(b):

1. $\mathcal{U}([0.45 \text{ m}, 0.55 \text{ m}] \times [-0.05 \text{ m}, 0.05 \text{ m}])$ (training distribution)

2. $\mathcal{U}([0.40\text{ m}, 0.60\text{ m}] \times [-0.10\text{ m}, 0.10\text{ m}])$
3. $\mathcal{U}([0.35\text{ m}, 0.65\text{ m}] \times [-0.15\text{ m}, 0.15\text{ m}])$
4. $\mathcal{U}([0.30\text{ m}, 0.70\text{ m}] \times [-0.20\text{ m}, 0.20\text{ m}])$
5. $\mathcal{U}([0.25\text{ m}, 0.75\text{ m}] \times [-0.25\text{ m}, 0.25\text{ m}])$
6. $\mathcal{U}([0.20\text{ m}, 0.80\text{ m}] \times [-0.30\text{ m}, 0.30\text{ m}])$

Control policy architecture. The manipulator's control policy architecture is provided in Fig. 4. The weights of the DNN are borrowed from [42] and the training is warm-started with the posterior provided in [42, App. A5.1].

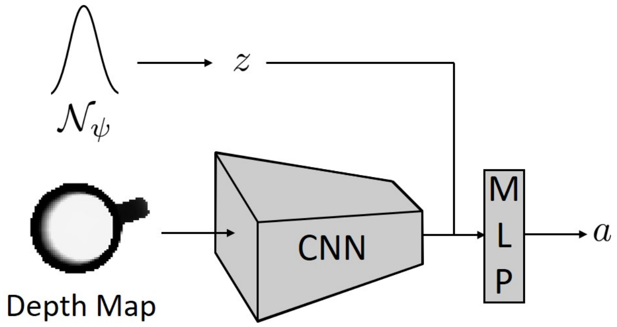

Figure 4: Network architecture for the manipulator's grasping control policy.

A.6.2 Vision-based obstacle avoidance with a drone

The approximate $C_{\mathcal{D}}(\pi)$ (estimated with $50,000$ held-out environments) is 0.149; PAC-Bayes thus provides a strong bound.

Environment generation. Training environments have 9 obstacles and have at least one gap which is wide enough to navigate through. We generate environments by randomly placing a set of cylindrical obstacles whose locations are sampled from the uniform distribution $\mathcal{U}([4.5\text{ m}, 7\text{ m}] \times [-3.5\text{ m}, 3.5\text{ m}])$ relative to the drone's starting point.

Training the prior. Training takes place completely in simulation. To allow for accurate sim-to-real transfer, the motion primitives are recorded trajectories of open-loop control inputs for the Parrot Swing hardware platform. We record multiple rollouts of each open-loop control policy. In

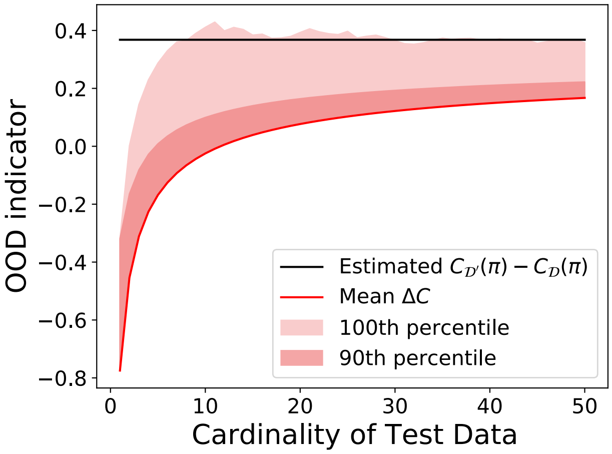

Figure 5: Numerical validation of lower bound in Theorem 3.

simulation, when the policy selects a motion primitive, we randomly select one of the corresponding recorded trajectories to run. We train the prior $\mathcal{N}_{\psi_0}$ over policies by transforming the problem into a supervised learning setting. For each of $10{,}000$ training environment in S the policy receives a depth map. Leveraging the simulation, we simulate each primitive (sampled uniformly from the set of recorded trajectories for that primitive) through each environment. We generate a label for each depth map by recording the minimum distance to an obstacle achieved by each of the primitives and passing the vector of distances through a softmax transformation. Note that even in simulation, we do not assume knowledge of the exact location of obstacles and record the closest distance as viewed by the robot's 120° field of view depth sensor. These depth maps and softmax labels can then be used for training the prior over policies in a supervised learning setting. We use the cross-entropy loss to train . The result is a policy trained to assign larger values to motion primitives which achieve a larger distance from obstacles.

Training platform. Training was performed on a desktop computer with an `Intel i7-8700k CPU` (12 CPU threads) and an `NVIDIA Titan Xp GPU` with 32 GB RAM.

Numerical validation of Theorem 3. We numerically validate our confidence bound in Fig. 5. We plot (i) the difference $C_{\mathcal{D}'}(\pi)$ - $C_{\mathcal{D}}(\pi)$ (estimated via exhaustive sampling of environments), (ii) the maximum computed lower-bound on $C_{\mathcal{D}'}(\pi)$ - $C_{\mathcal{D}}(\pi)$ (computed using a confidence level of 0.9) over $500{,}000$ datasets S', and (iii) the $90\mathrm{th}$ percentile value of the bound over the $100{,}000$ datasets. As guaranteed by Theorem 3, the bound is valid greater than 90% of the time.