APPENDIX A
SHIFT OPERATOR

We discuss some details in defining the shift operator. Let $\boldsymbol{\theta}_{t-1}$ be the approximate solution to the previous problem and $\tilde{\boldsymbol{\theta}}_t$ denote the initial condition of $\boldsymbol{\theta}$ in solving (3), and consider sampling $\hat{\boldsymbol{u}}_t \sim \boldsymbol{\pi}_{\tilde{\boldsymbol{\theta}}_t}$ and $\hat{\boldsymbol{u}}_{t-1} \sim \boldsymbol{\pi}_{\boldsymbol{\theta}_{t-1}}$. We set

$$\tilde{\boldsymbol{\theta}}_t = \Phi(\boldsymbol{\theta}_{t-1})$$

by defining a *shift operator* Φ that outputs a new parameter in $\boldsymbol{\Theta}$. This Φ can be chosen to satisfy desired properties, one example being that when conditioned on $\hat{u}_{t-1}$ and x_t, the marginal distributions of $\hat{u}_t, \ldots, \hat{u}_{t+H-2}$ are the same for both $\hat{\boldsymbol{u}}_t$ of $\boldsymbol{\pi}_{\tilde{\boldsymbol{\theta}}_t}$ and $\hat{\boldsymbol{u}}_{t-1}$ of $\boldsymbol{\pi}_{\boldsymbol{\theta}_{t-1}}$. A simple example of this property is shown in Fig. 8. Note that $\hat{\boldsymbol{u}}_t$ also involves a new control $\hat{u}_{t+H-1}$ that is not in $\hat{\boldsymbol{u}}_{t-1}$, so the choice of Φ is not unique but algorithm dependent; for example, we can set $\hat{u}_{t+H-1}$ of $\boldsymbol{\pi}_{\tilde{\boldsymbol{\theta}}_t}$ to follow the same distribution as $\hat{u}_{t+H-2}$ (cf. Section III-B). Because the subproblems in (3) of two consecutive time steps share all control variables except for the first and the last ones, the "shifted" parameter $\Phi(\boldsymbol{\theta}_{t-1})$ to the current problem should be almost as good as the optimized parameter $\boldsymbol{\theta}_{t-1}$ is to the previous problem. In other words, setting $\tilde{\boldsymbol{\theta}}_t = \Phi(\boldsymbol{\theta}_{t-1})$ provides a warm start to (3) and amortizes the computational complexity of solving for $\boldsymbol{\theta}_t$.

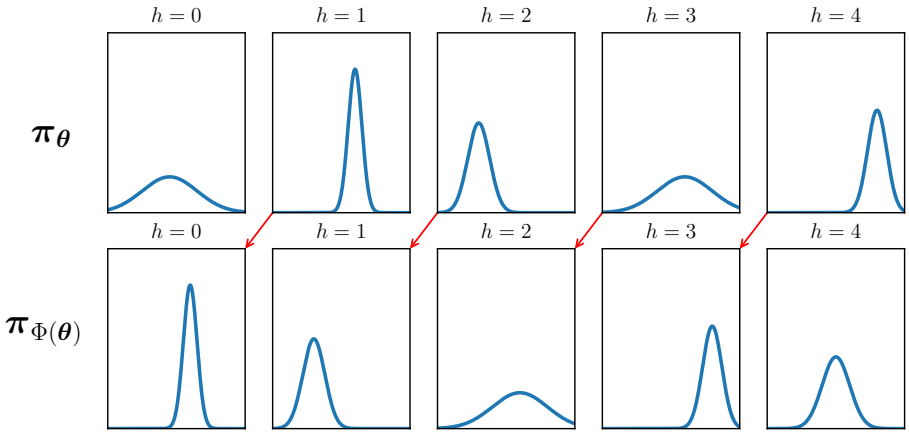


Fig. 8: A simple example of the shift operator Φ. Here, the control distribution $\boldsymbol{\pi}_{\boldsymbol{\theta}}$ consists of a sequence of $H = 5$ independent Gaussian distributions. The shift operator moves the parameters of the Gaussians one time step forward and replaces the parameters at $h = 4$ with some default parameters.

APPENDIX B
VARIATIONS OF DMD-MPC

The control distributions in DMD-MPC can be fairly general (in addition to the categorical and Gaussian distributions that we discussed) and control constraints on the problem (e.g., control limits) can be directly incorporated through proper choices of control distributions, such as the beta distribution, or through mapping the unconstrained control through some squashing function (e.g., $\tanh$ or clamp). Though our framework cannot directly handle state constraints as in constrained optimization approaches, a constraint can be relaxed to an indicator function which activates if the constraint is violated. The indicator function can then be added to the cost function in (4) with some weight that encodes how strictly the constraint should be enforced.

Moreover, different integration techniques, such as Gaussian quadrature [5], can be adopted to replace the likelihood-ratio derivative in (9) for computing the required gradient direction. We also note that the independence assumption on the control distribution in (18) is not necessary in our framework; time-correlated control distributions and feedback policies are straightforward to consider in DMD-MPC.

APPENDIX C
PROOFS

Proof of Proposition 1: We prove the first statement; the second one follows directly from the duality relationship. The statement follows from the derivations below; we can write

$$\begin{aligned}
\eta_{t+1} &= \arg\min_{\eta\in\mathcal{H}} \langle \gamma_t g_t, \eta\rangle + D_A(\eta\|\eta_t) \\
&= \arg\min_{\eta\in\mathcal{H}} \langle \gamma_t g_t, \eta\rangle + A(\eta) - \langle \nabla A(\eta_t), \eta\rangle \\
&= \arg\min_{\eta\in\mathcal{H}} \langle \gamma_t g_t - \mu_t, \eta\rangle + A(\eta) \\
&= \arg\max_{\eta\in\mathcal{H}} \langle \mu_t - \gamma_t g_t, \eta\rangle - A(\eta) \\
&= \nabla A^*(\mu_t - \gamma_t g_t)
\end{aligned}$$

where the last equality is due to the assumption that $\mu_t - \gamma_t g_t \in \mathcal{M}$. Then applying ∇A on both sides and using the relationship that $\nabla A = (\nabla A^*)^{-1}$, we have $\mu_{t+1} = \nabla A(\eta_{t+1}) = \mu_t - \gamma_t g_t$. ■

APPENDIX D
DERIVATION OF LQR AND LEQR LOSSES

The dynamics in Equation (1) are given by

$$x_{t+1} = Ax_t + Bu_t + w_t$$

for some matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ and $w_t \sim \mathcal{N}(0, W)$, where $W \in \mathbb{S}^n_{++}$. For a control sequence $\hat{\boldsymbol{u}}_t$, noise sequence $\hat{\boldsymbol{w}}_t$, and initial state x_t, the resulting state sequence $\hat{\boldsymbol{x}}_t$ is found through convolution:

$$\begin{bmatrix} \hat{x}_t \\ \hat{x}_{t+1} \\ \hat{x}_{t+2} \\ \vdots \\ \hat{x}_{t+H} \end{bmatrix} = \begin{bmatrix} I \\ A \\ A^2 \\ \vdots \\ A^H \end{bmatrix} x_t + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ B & 0 & \cdots & 0 \\ AB & B & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A^{H-1}B & A^{H-2}B & \cdots & B \end{bmatrix} \begin{bmatrix} \hat{u}_t \\ \hat{u}_{t+1} \\ \vdots \\ \hat{u}_{t+H-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & \cdots & 0 \\ I & 0 & \cdots & 0 \\ A & I & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A^{H-1} & A^{H-2} & \cdots & I \end{bmatrix} \begin{bmatrix} \hat{w}_t \\ \hat{w}_{t+1} \\ \vdots \\ \hat{w}_{t+H-1} \end{bmatrix},$$

or, in matrix form:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{F}x_t + \boldsymbol{G}\hat{\boldsymbol{u}}_t + \boldsymbol{L}\hat{\boldsymbol{w}}_t,$$

where $\boldsymbol{F}$, $\boldsymbol{G}$, and $\boldsymbol{L}$ are defined naturally from the convolution equation above. Note that $\hat{\boldsymbol{w}}_t \sim \mathcal{N}(0, \boldsymbol{W})$, where $\boldsymbol{W} = \mathrm{diag}(W, W, \ldots, W, W)$. Thus, we also have that

$$\hat{\boldsymbol{x}}_t \sim \mathcal{N}(\boldsymbol{F}x_t + \boldsymbol{G}\hat{\boldsymbol{u}}_t, \boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}).$$

We define the instantaneous and terminal costs as

$$\begin{aligned}
c(x, u) &= \frac{1}{2}x^\mathsf{T}Qx + \frac{1}{2}u^\mathsf{T}Ru \\
c_{\mathrm{end}}(x) &= \frac{1}{2}x^\mathsf{T}Q_{\mathrm{end}}x,
\end{aligned}$$

where $Q, Q_{\mathrm{end}} \in \mathbb{S}^n_+$ and $R \in \mathbb{S}^m_{++}$. Thus, the statistic $C(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{u}}_t)$ is

$$C(\hat{\boldsymbol{x}}_t, \hat{\boldsymbol{u}}_t) = \frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{x}}_t + \frac{1}{2}\hat{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{R}\hat{\boldsymbol{u}}_t,$$

where $\boldsymbol{Q} = \mathrm{diag}(Q, Q, \ldots, Q, Q_{\mathrm{end}})$ and $\boldsymbol{R} = \mathrm{diag}(R, R, \ldots, R, R)$.

Our control distribution is a Dirac delta distribution located at the given parameter: $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\hat{\boldsymbol{u}}_t) = \delta(\hat{\boldsymbol{u}}_t - \boldsymbol{\theta})$.

A. LQR

The loss is defined as $\ell_t(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\pi_\theta}, \hat{\boldsymbol{x}}_t}\left[\frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{x}}_t + \frac{1}{2}\hat{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{R}\hat{\boldsymbol{u}}_t\right]$. Expanding this out gives:

$$\begin{aligned}
\ell_t(\boldsymbol{\theta}) &= \mathbb{E}_{\boldsymbol{\pi_\theta}, \hat{\boldsymbol{x}}_t}\left[\frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{x}}_t + \frac{1}{2}\hat{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{R}\hat{\boldsymbol{u}}_t\right] \\
&= \frac{1}{2}\boldsymbol{\theta}^\mathsf{T}\big(\boldsymbol{G}^\mathsf{T}\boldsymbol{Q}\boldsymbol{G} + \boldsymbol{R}\big)\boldsymbol{\theta} + x_t^\mathsf{T}\boldsymbol{F}^\mathsf{T}\boldsymbol{Q}\boldsymbol{G}\boldsymbol{\theta} + \frac{1}{2}x_t^\mathsf{T}\boldsymbol{F}^\mathsf{T}\boldsymbol{Q}\boldsymbol{F}x_t + \frac{1}{2}\mathbb{E}\big[\hat{\boldsymbol{w}}_t^\mathsf{T}\boldsymbol{L}^\mathsf{T}\boldsymbol{Q}\boldsymbol{L}\hat{\boldsymbol{w}}_t\big] \\
&= \frac{1}{2}\boldsymbol{\theta}^\mathsf{T}\big(\boldsymbol{G}^\mathsf{T}\boldsymbol{Q}\boldsymbol{G} + \boldsymbol{R}\big)\boldsymbol{\theta} + x_t^\mathsf{T}\boldsymbol{F}^\mathsf{T}\boldsymbol{Q}\boldsymbol{G}\boldsymbol{\theta} + \frac{1}{2}x_t^\mathsf{T}\boldsymbol{F}^\mathsf{T}\boldsymbol{Q}\boldsymbol{F}x_t + \frac{1}{2}\mathrm{tr}\big(\boldsymbol{Q}\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\big).
\end{aligned}$$

We see this is a quadratic problem in $\boldsymbol{\theta}$ by defining

$$\begin{aligned}
\boldsymbol{R}_t &= \boldsymbol{G}^\mathsf{T}\boldsymbol{Q}\boldsymbol{G} + \boldsymbol{R} \\
\boldsymbol{r}_t &= \boldsymbol{G}^\mathsf{T}\boldsymbol{Q}\boldsymbol{F}x_t.
\end{aligned}$$

B. LEQR

The loss is defined as

$$\ell_t(\boldsymbol{\theta}) = -\log \mathbb{E}_{\boldsymbol{\pi_\theta}, \hat{\boldsymbol{x}}_t}\left[\exp\left(-\frac{1}{\lambda}\left(\frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}\hat{\boldsymbol{x}}_t + \frac{1}{2}\hat{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{R}\hat{\boldsymbol{u}}_t\right)\right)\right]$$

for some parameter $\lambda > 0$. For compactness, we define $\boldsymbol{Q}' = \frac{1}{\lambda}\boldsymbol{Q}$ and $\boldsymbol{R}' = \frac{1}{\lambda}\boldsymbol{R}$ so that the exponent contains $-\frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}'\hat{\boldsymbol{x}} - \frac{1}{2}\hat{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{R}'\hat{\boldsymbol{u}}_t$. In expanding the loss, we use the following fact:

Fact 2. *For $x \sim \mathcal{N}(\mu, \Sigma)$, where $\Sigma \in \mathbb{S}_{++}^n$, and constants $A \in \mathbb{S}_+^n$ and $b \in \mathbb{R}^n$:*

$$\mathbb{E}_x\left[\exp\left(-\frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x\right)\right] = \frac{1}{\sqrt{|A\Sigma + I|}}\exp\left(-\frac{1}{2}\big(\mu^\mathsf{T}\Sigma^{-1}\mu - (\Sigma^{-1}\mu - b)^\mathsf{T}(A + \Sigma^{-1})^{-1}(\Sigma^{-1}\mu - b)\big)\right).$$

Proof: We expand the expectation and complete the square:

$$\begin{aligned}
\mathbb{E}_x\left[\exp\left(-\frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x\right)\right] &= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int \exp\left(-\frac{1}{2}(x-\mu)^\mathsf{T}\Sigma^{-1}(x-\mu)\right)\exp\left(-\frac{1}{2}x^\mathsf{T}Ax - b^\mathsf{T}x\right)\mathrm{d}x \\
&= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\int \exp\left(-\frac{1}{2}\big[x^\mathsf{T}(A+\Sigma^{-1})x + 2(b - \Sigma^{-1}\mu)^\mathsf{T}x + \mu^\mathsf{T}\Sigma^{-1}\mu\big]\right)\mathrm{d}x \\
&= \frac{1}{\sqrt{(2\pi)^n|\Sigma|}}\exp(c)\int \exp\left(-\frac{1}{2}(x-\tilde{\mu})^\mathsf{T}\tilde{\Sigma}^{-1}(x-\tilde{\mu})\right)\mathrm{d}x \\
&= \frac{\sqrt{(2\pi)^n|\tilde{\Sigma}|}}{\sqrt{(2\pi)^n|\Sigma|}}\exp(c) \\
&= \frac{1}{\sqrt{|A+\Sigma^{-1}||\Sigma|}}\exp(c) \\
&= \frac{1}{\sqrt{|A\Sigma + I|}}\exp\left(-\frac{1}{2}\big(\mu^\mathsf{T}\Sigma^{-1}\mu - (\Sigma^{-1}\mu - b)^\mathsf{T}(A+\Sigma^{-1})^{-1}(\Sigma^{-1}\mu - b)\big)\right),
\end{aligned}$$

where $\tilde{\mu} = (A+\Sigma^{-1})^{-1}(\Sigma^{-1}\mu - b)$, $\tilde{\Sigma} = (A+\Sigma^{-1})^{-1}$, and $c = -\frac{1}{2}\big(\mu^\mathsf{T}\Sigma^{-1}\mu - (\Sigma^{-1}\mu - b)^\mathsf{T}(A+\Sigma^{-1})^{-1}(\Sigma^{-1}\mu - b)\big)$. ■

We now expand the loss:

$$\begin{aligned}
\ell_t(\boldsymbol{\theta}) &= -\log \mathbb{E}_{\boldsymbol{\pi}_{\boldsymbol{\theta}}, \hat{\boldsymbol{x}}_t}\left[\exp\left(-\frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}'\hat{\boldsymbol{x}}_t - \frac{1}{2}\hat{\boldsymbol{u}}_t^\mathsf{T}\boldsymbol{R}'\hat{\boldsymbol{u}}_t\right)\right] \\
&= -\log \mathbb{E}_{\hat{\boldsymbol{x}}_t}\left[\exp\left(-\frac{1}{2}\hat{\boldsymbol{x}}_t^\mathsf{T}\boldsymbol{Q}'\hat{\boldsymbol{x}}_t - \frac{1}{2}\boldsymbol{\theta}^\mathsf{T}\boldsymbol{R}'\boldsymbol{\theta}\right)\right] \\
&= -\log\left\{\frac{1}{\sqrt{|\boldsymbol{Q}'\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}+I|}}\exp\Big(-\frac{1}{2}\big[(\boldsymbol{F}x_t+\boldsymbol{G}\boldsymbol{\theta})^\mathsf{T}(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T})^{-1}(\boldsymbol{F}x_t+\boldsymbol{G}\boldsymbol{\theta})\right. \right. \\
&\qquad\qquad - (\boldsymbol{F}x_t+\boldsymbol{G}\boldsymbol{\theta})^\mathsf{T}(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\boldsymbol{Q}'\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}+\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T})^{-1}(\boldsymbol{F}x_t+\boldsymbol{G}\boldsymbol{\theta}) \\
&\qquad\qquad \left.\left. + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{R}'\boldsymbol{\theta}\big]\Big)\right\} \\
&= \frac{1}{2}\big[(\boldsymbol{F}x_t+\boldsymbol{G}\boldsymbol{\theta})^\mathsf{T}[(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T})^{-1}+(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\boldsymbol{Q}'\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}+\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T})^{-1}](\boldsymbol{F}x_t+\boldsymbol{G}\boldsymbol{\theta}) + \boldsymbol{\theta}^\mathsf{T}\boldsymbol{R}'\boldsymbol{\theta}\big] + \frac{1}{2}\log|\boldsymbol{Q}'\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}+I|.
\end{aligned}$$

We see this is a quadratic problem in $\boldsymbol{\theta}$ by defining

$$\begin{aligned}
\boldsymbol{R}_t &= \boldsymbol{G}^\mathsf{T}\left[(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T})^{-1}+\left(\frac{1}{\lambda}\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\boldsymbol{Q}\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}+\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\right)^{-1}\right]\boldsymbol{G}+\frac{1}{\lambda}\boldsymbol{R} \\
\boldsymbol{r}_t &= \boldsymbol{G}^\mathsf{T}\left[(\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T})^{-1}+\left(\frac{1}{\lambda}\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\boldsymbol{Q}\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}+\boldsymbol{L}\boldsymbol{W}\boldsymbol{L}^\mathsf{T}\right)^{-1}\right]\boldsymbol{F}x_t.
\end{aligned}$$

APPENDIX E
EXPERIMENTAL SETUP

A. Cartpole

The state is $x_t = (p_t, \varphi_t, v_t, \dot{\varphi}_t)$, where p_t is the cart position, φ_t is the pole's angle, v_t and $\dot{\varphi}_t$ are the corresponding velocities, and the control u_t is the force applied to the cart. We define the instantaneous cost and terminal cost of the MPC problem as

$$\begin{aligned}
c(x_t, u_t) &= 10p_t^2 + 500(\varphi_t - \pi)^2 + v_t^2 + 15\dot{\varphi}_t^2 + 1000 \cdot \mathbf{1}\{|\varphi_t - \pi| \geq \Delta\} \\
c_{\text{end}}(x_t) &= c(x_t, 0)
\end{aligned}$$

where Δ is some threshold. For our experiments, we set $\Delta = 12^\circ = 0.21$ radians.

In our experiments, the pole is massless except for some weight at the end of the pole. The mass of the cart and pole weight are 0.711 kg and 0.209 kg, respectively. The true length of the pole is 0.326 m, whereas the length used in the model is 0.346 m. Each time step is modeled using an Euler discretization of 0.02 seconds. Each episode of the problem lasts 500 time steps (i.e, 10 seconds) and has episode cost equal to the sum of encountered instantaneous costs. Both the true system and the model apply Gaussian additive noise to the commanded control with zero mean and a standard deviation of 5 newtons. For the continuous system, the commanded control is clamped to ± 25 newtons. For the discrete system, the controller can either command 10 newtons to the left, 10 newtons to the right, or 0 newtons.

Both the discrete and continuous controller use a planning horizon of 50 time steps (i.e., 1 second). For the continuous controller, we keep the standard deviation of the Gaussian distribution fixed at 2 newtons for each time step in the planning horizon. When applying a control u_t on the real cartpole, we choose the mode of π_{θ_t} rather than sample from the distribution.

All reported results were gathered using ten episodes per parameter setting.

B. AutoRally

The state of the vehicle is $x_t = (p_{x,t}, p_{y,t}, \varphi_t, r_t, v_{x,t}, v_{y,t}, \dot{\varphi}_t)$, where $(p_{x,t}, p_{y,t})$ is the position of the car in the global frame, φ_t and r_t are the yaw and roll angles, $v_{x,t}$ and $v_{y,t}$ are the longitudinal and lateral velocities in the car frame, and $\dot{\varphi}_t$ is the yaw rate. The control u_t we apply is the throttle and steering angle. For some weights $w_1, \ldots, w_4$, the cost function is

$$\begin{aligned}
c(x_t, u_t) &= w_1|s_t - s_{\text{tgt}}|^k + w_2 M(p_{x,t}, p_{y,t}) + w_3 S_c(x_t) \\
c_{\text{end}}(\boldsymbol{x}_t) &= w_4 C(\boldsymbol{x}_t).
\end{aligned}$$

Here, s_t and s_{tgt} are the current and target speed of the car, respectively. Note the speed is calculated as $s_t = \sqrt{v_{x,t}^2 + v_{y,t}^2}$. $M(p_{x,t}, p_{y,t})$ is the positional cost of the car (low cost in center of track, high cost at edge of track), $S_c(x_t)$ is an indicator

variable which activates if the slip angle[16] exceeds a certain threshold, and $C(\boldsymbol{x}_t)$ is an indicator function which activates if the car leaves the track at all in the trajectory. Note that the terminal cost depends on the trajectory instead of the terminal state. Each time step represents 0.02 seconds and the length of the planning trajectory is 100 time steps (i.e., 2 seconds). The values for the cost function parameters are given in Table II.

The control space for each of the throttle and steering angle is normalized to the range $[-1, 1]$. For our experiments, we clamp the throttle to $[-1, 0.65]$. In simulated experiments, the standard deviations of the throttle and steering angle distributions were 0.3 and 0.275, respectively. In the real world experiments, they were both set to 0.3. When applying a control u_t on the car, we chose the mean of π_{θ_t} rather than sampling from the distribution.

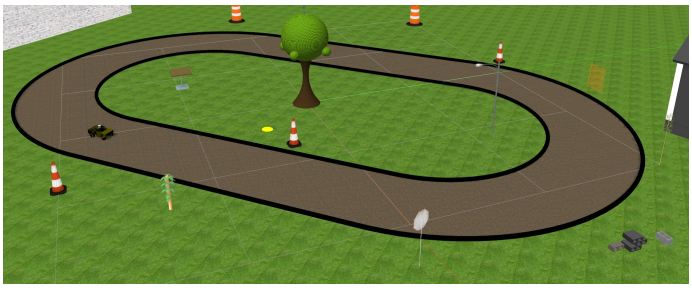


Fig. 9: Simulated AutoRally task.

In simulation, the environment (Fig. 9) is an elliptical track approximately 3 meters wide and 30 meters across at its furthest point. The real-world dirt track is about 5 meters wide and and has a track length of 170 meters. All reported results for simulated experiments were gathered using 30 consecutive laps in the counter-clockwise direction for each parameter setting, whereas for real-world experiments results were gathered using ten laps for each parameter setting.

TABLE II: Cost function settings for AutoRally experiments.

	s_{tgt} (m/s)	k	w_1	w_2	w_3	w_4	Slip angle threshold (rad)
Gazebo simulator	11	1	30	250	10	10000	0.275
Real world	9	2	4.25	200	100	10000	0.9

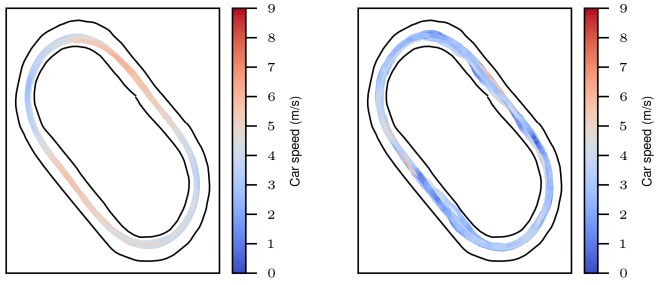
APPENDIX F
EXTRA EXPERIMENTAL RESULTS

A. Simulated Experiments

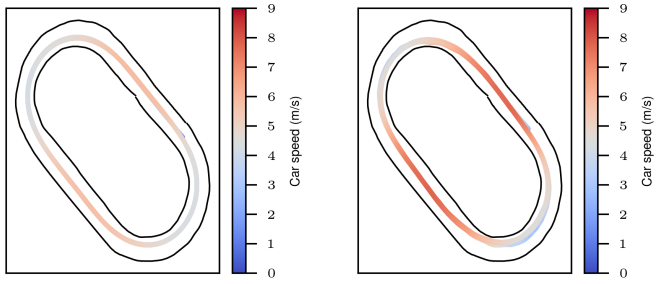
Adding onto the results from Section V-B2, we qualitatively evaluate two particular extremes: few vs. many samples (64 vs. 3840) and small vs. large step size (0.5 vs. 1) by looking at the path and speed of the car during the episode (Fig. 10). At small step sizes (Figs. 10a and 10c), the path and speed profiles are rather similar, while with few samples and a large step size (Fig. 10b), the car drives much more slowly and erratically, sometimes even stopping. In the ideal scenario with many samples and a large step size, the car can achieve consistently high speed while driving smoothly (Fig. 10d).

We also experimented with instead optimizing the expected cost (10) and found performance was dramatically worse (Fig. 11), even when using 3840 samples per gradient. At best, the car would drive in the center of the track at speeds below 4 m/s (Fig. 11c), and at worst, the car would either slowly drive along the track walls (Fig. 11a) or the controller would eventually produce NaN controls that would prematurely end the experiment (Fig. 11d). This poor performance is likely due to most samples in the estimate of (11) having very high cost (e.g., due to leaving the track) and contributing significantly to the gradient estimate. On the other hand, when estimating (17), as in the experiments in Section V-B2, these high cost trajectories are assigned very low weights so that only low cost trajectories contribute to the gradient estimate.

[16]The slip angle is defined as $-\arctan\frac{v_{y,t}}{|v_{x,t}|}$, which gives the angle between the direction the car is pointing and the direction in which it is actually traveling.

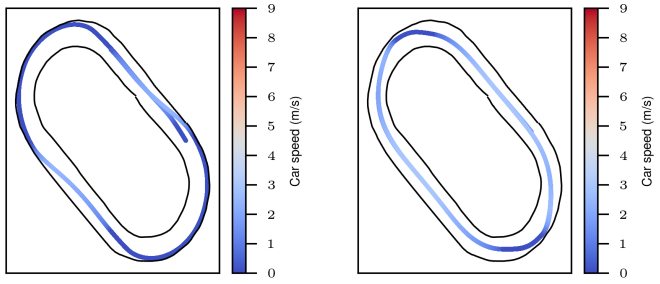


(a) 64 samples, $\gamma_t = 0.5$ (b) 64 samples, $\gamma_t = 1$

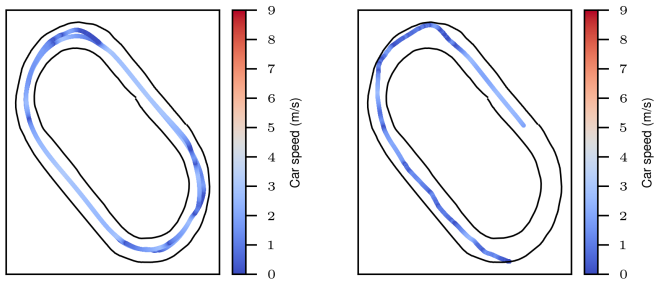


(c) 3840 samples, $\gamma_t = 0.5$ (d) 3840 samples, $\gamma_t = 1$

Fig. 10: Car speeds when optimizing the exponential utility (16). The speeds and trajectories are very similar at step size 0.5, irrespective of the number of samples. At step size 1, though, 64 samples result in capricious maneuvers and low speeds, whereas 3840 samples result in smooth driving at high speeds.



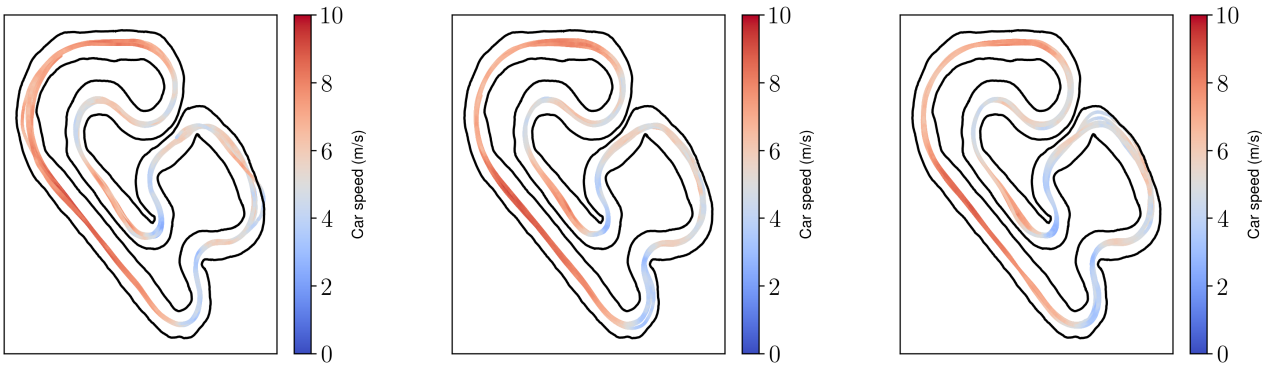
(a) $\gamma_t = 0.025$ (b) $\gamma_t = 0.05$



(c) $\gamma_t = 0.075$ (d) $\gamma_t = 0.1$

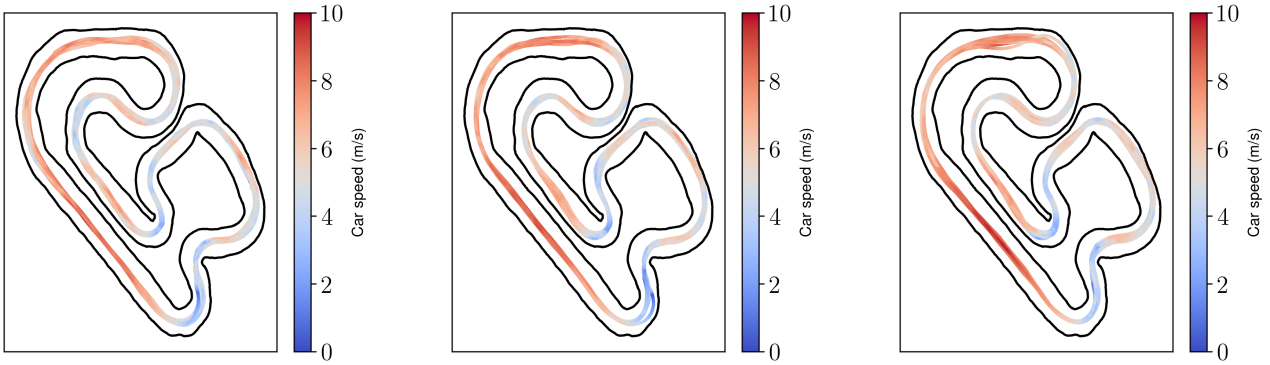
Fig. 11: Car speeds when optimizing the expected cost (10). All tested step sizes result in low speeds. At too low or too high or a step size, the car will drive along the wall or crash into it.

B. Figures for Real-World Experiments



(a) $\gamma_t = 1$ (b) $\gamma_t = 0.8$ (c) $\gamma_t = 0.6$

Fig. 12: Car speeds with 1920 samples per gradient estimate.



(a) $\gamma_t = 1$ (b) $\gamma_t = 0.8$ (c) $\gamma_t = 0.6$

Fig. 13: Car speeds with 64 samples per gradient estimate.