10. Appendix

10.1. Proof of Policy Performance Bound

10.1.1. Preliminaries

Our analysis will make extensive use of the discounted future state distribution, d^{π}, which is defined as

$$d^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|\pi).$$

It allows us to express the expected discounted total reward compactly as

$$J(\pi) = \frac{1}{1-\gamma}\mathop{\mathrm{E}}_{\substack{s\sim d^{\pi}\\ a\sim\pi\\ s'\sim P}}\left[R(s,a,s')\right], \tag{17}$$

where by $a \sim \pi$, we mean $a \sim \pi(\cdot|s)$, and by $s' \sim P$, we mean $s' \sim P(\cdot|s,a)$. We drop the explicit notation for the sake of reducing clutter, but it should be clear from context that a and s' depend on s.

First, we examine some useful properties of d^{π} that become apparent in vector form for finite state spaces. Let $p_{\pi}^t \in \mathbb{R}^{|S|}$ denote the vector with components $p_{\pi}^t(s) = P(s_t = s|\pi)$, and let $P_{\pi} \in \mathbb{R}^{|S|\times|S|}$ denote the transition matrix with components $P_{\pi}(s'|s) = \int da P(s'|s,a)\pi(a|s)$; then $p_{\pi}^t = P_{\pi}p_{\pi}^{t-1} = P_{\pi}^t\mu$ and

$$\begin{aligned} d^{\pi} &= (1-\gamma)\sum_{t=0}^{\infty}(\gamma P_{\pi})^t\mu \\ &= (1-\gamma)(I-\gamma P_{\pi})^{-1}\mu. \end{aligned} \tag{18}$$

This formulation helps us easily obtain the following lemma.

Lemma 1. *For any function $f : S \to \mathbb{R}$ and any policy π,*

$$(1-\gamma)\mathop{\mathrm{E}}_{s\sim\mu}\left[f(s)\right] + \mathop{\mathrm{E}}_{\substack{s\sim d^{\pi}\\ a\sim\pi\\ s'\sim P}}\left[\gamma f(s')\right] - \mathop{\mathrm{E}}_{s\sim d^{\pi}}\left[f(s)\right] = 0. \tag{19}$$

Proof. Multiply both sides of (18) by $(I-\gamma P_{\pi})$ and take the inner product with the vector $f \in \mathbb{R}^{|S|}$. □

Combining this with (17), we obtain the following, for any function f and any policy π:

$$J(\pi) = \mathop{\mathrm{E}}_{s\sim\mu}\left[f(s)\right] + \frac{1}{1-\gamma}\mathop{\mathrm{E}}_{\substack{s\sim d^{\pi}\\ a\sim\pi\\ s'\sim P}}\left[R(s,a,s') + \gamma f(s') - f(s)\right]. \tag{20}$$

This identity is nice for two reasons. First: if we pick f to be an approximator of the value function V^{π}, then (20) relates the true discounted return of the policy ($J(\pi)$) to the estimate of the policy return ($\mathrm{E}_{s\sim\mu}[f(s)]$) and to the on-policy average TD-error of the approximator; this is aesthetically satisfying. Second: it shows that reward-shaping by $\gamma f(s') - f(s)$ has the effect of translating the total discounted return by $\mathrm{E}_{s\sim\mu}[f(s)]$, a fixed constant independent of policy; this illustrates the finding of Ng. et al. (1999) that reward shaping by $\gamma f(s') + f(s)$ does not change the optimal policy.

It is also helpful to introduce an identity for the vector difference of the discounted future state visitation distributions on two different policies, π' and π. Define the matrices $G \doteq (I-\gamma P_{\pi})^{-1}$, $\bar{G} \doteq (I-\gamma P_{\pi'})^{-1}$, and $\Delta = P_{\pi'} - P_{\pi}$. Then:

$$\begin{aligned} G^{-1} - \bar{G}^{-1} &= (I-\gamma P_{\pi}) - (I-\gamma P_{\pi'}) \\ &= \gamma\Delta; \end{aligned}$$

left-multiplying by G and right-multiplying by $\bar{G}$, we obtain

$$\bar{G} - G = \gamma\bar{G}\Delta G.$$

Thus

$$\begin{aligned} d^{\pi'} - d^{\pi} &= (1-\gamma)\left(\bar{G} - G\right)\mu \\ &= \gamma(1-\gamma)\bar{G}\Delta G\mu \\ &= \gamma\bar{G}\Delta d^{\pi}. \end{aligned} \tag{21}$$

For simplicity in what follows, we will only consider MDPs with finite state and action spaces, although our attention is on MDPs that are too large for tabular methods.

10.1.2. Main Results

In this section, we will derive and present the new policy improvement bound. We will begin with a lemma:

Lemma 2. *For any function $f : S \to \mathbb{R}$ and any policies π' and π, define*

$$L_{\pi,f}(\pi') \doteq \underset{\substack{s\sim d^{\pi} \\ a\sim\pi \\ s'\sim P}}{\mathrm{E}}\left[\left(\frac{\pi'(a|s)}{\pi(a|s)} - 1\right)\left(R(s,a,s') + \gamma f(s') - f(s)\right)\right], \tag{22}$$

and $\epsilon_f^{\pi'} \doteq \max_s \left|\mathrm{E}_{a\sim\pi',s'\sim P}[R(s,a,s') + \gamma f(s') - f(s)]\right|$. Then the following bounds hold:

$$J(\pi') - J(\pi) \geq \frac{1}{1-\gamma}\left(L_{\pi,f}(\pi') - 2\epsilon_f^{\pi'} D_{TV}(d^{\pi'}||d^{\pi})\right), \tag{23}$$

$$J(\pi') - J(\pi) \leq \frac{1}{1-\gamma}\left(L_{\pi,f}(\pi') + 2\epsilon_f^{\pi'} D_{TV}(d^{\pi'}||d^{\pi})\right), \tag{24}$$

where D_{TV} is the total variational divergence. Furthermore, the bounds are tight (when $\pi' = \pi$, the LHS and RHS are identically zero).

Proof. First, for notational convenience, let $\delta_f(s,a,s') \doteq R(s,a,s') + \gamma f(s') - f(s)$. (The choice of δ to denote this quantity is intentionally suggestive—this bears a strong resemblance to a TD-error.) By (20), we obtain the identity

$$J(\pi') - J(\pi) = \frac{1}{1-\gamma}\left(\underset{\substack{s\sim d^{\pi'} \\ a\sim\pi' \\ s'\sim P}}{\mathrm{E}}\left[\delta_f(s,a,s')\right] - \underset{\substack{s\sim d^{\pi} \\ a\sim\pi \\ s'\sim P}}{\mathrm{E}}\left[\delta_f(s,a,s')\right].\right)$$

Now, we restrict our attention to the first term in this equation. Let $\bar{\delta}_f^{\pi'} \in \mathbb{R}^{|S|}$ denote the vector of components $\bar{\delta}_f^{\pi'}(s) = \mathrm{E}_{a\sim\pi',s'\sim P}[\delta_f(s,a,s')|s]$. Observe that

$$\begin{aligned} \underset{\substack{s\sim d^{\pi'} \\ a\sim\pi' \\ s'\sim P}}{\mathrm{E}}\left[\delta_f(s,a,s')\right] &= \left\langle d^{\pi'}, \bar{\delta}_f^{\pi'}\right\rangle \\ &= \left\langle d^{\pi}, \bar{\delta}_f^{\pi'}\right\rangle + \left\langle d^{\pi'} - d^{\pi}, \bar{\delta}_f^{\pi'}\right\rangle \end{aligned}$$

This term is then straightforwardly bounded by applying Hölder's inequality; for any $p, q \in [1,\infty]$ such that $1/p + 1/q = 1$, we have

$$\left\langle d^{\pi}, \bar{\delta}_f^{\pi'}\right\rangle + \left\|d^{\pi'} - d^{\pi}\right\|_p \left\|\bar{\delta}_f^{\pi'}\right\|_q \geq \underset{\substack{s\sim d^{\pi'} \\ a\sim\pi' \\ s'\sim P}}{\mathrm{E}}\left[\delta_f(s,a,s')\right] \geq \left\langle d^{\pi}, \bar{\delta}_f^{\pi'}\right\rangle - \left\|d^{\pi'} - d^{\pi}\right\|_p \left\|\bar{\delta}_f^{\pi'}\right\|_q.$$

The lower bound leads to (23), and the upper bound leads to (24).

We choose $p = 1$ and $q = \infty$; however, we believe that this step is very interesting, and different choices for dealing with the inner product $\left\langle d^{\pi'} - d^{\pi}, \bar{\delta}_f^{\pi'}\right\rangle$ may lead to novel and useful bounds.

With $\left\|d^{\pi'} - d^{\pi}\right\|_1 = 2D_{TV}(d^{\pi'}||d^{\pi})$ and $\left\|\bar{\delta}_f^{\pi'}\right\|_\infty = \epsilon_f^{\pi'}$, the bounds are almost obtained. The last step is to observe that, by the importance sampling identity,

$$\begin{aligned} \left\langle d^{\pi}, \bar{\delta}_f^{\pi'} \right\rangle &= \underset{\substack{s\sim d^{\pi} \\ a\sim\pi' \\ s'\sim P}}{\mathrm{E}} \left[\delta_f(s,a,s')\right] \\ &= \underset{\substack{s\sim d^{\pi} \\ a\sim\pi \\ s'\sim P}}{\mathrm{E}} \left[\left(\frac{\pi'(a|s)}{\pi(a|s)}\right)\delta_f(s,a,s')\right]. \end{aligned}$$

After grouping terms, the bounds are obtained. □

This lemma makes use of many ideas that have been explored before; for the special case of $f = V^{\pi}$, this strategy (after bounding $D_{TV}(d^{\pi'}||d^{\pi})$) leads directly to some of the policy improvement bounds previously obtained by Pirotta et al. and Schulman et al. The form given here is slightly more general, however, because it allows for freedom in choosing f.

Remark. It is reasonable to ask if there is a choice of f which maximizes the lower bound here. This turns out to trivially be $f = V^{\pi'}$. Observe that $\mathrm{E}_{s'\sim P}\left[\delta_{V^{\pi'}}(s,a,s')|s,a\right] = A^{\pi'}(s,a)$. For all states, $\mathrm{E}_{a\sim\pi'}[A^{\pi'}(s,a)] = 0$ (by the definition of $A^{\pi'}$), thus $\bar{\delta}^{\pi'}_{V^{\pi'}} = 0$ and $\epsilon^{\pi'}_{V^{\pi'}} = 0$. Also, $L_{\pi,V^{\pi'}}(\pi') = -\mathrm{E}_{s\sim d^{\pi},a\sim\pi}\left[A^{\pi'}(s,a)\right]$; from (20) with $f = V^{\pi'}$, we can see that this exactly equals $J(\pi') - J(\pi)$. Thus, for $f = V^{\pi'}$, we recover an exact equality. While this is not practically useful to us (because, when we want to optimize a lower bound with respect to π', it is too expensive to evaluate $V^{\pi'}$ for each candidate to be practical), it provides insight: the penalty coefficient on the divergence captures information about the mismatch between f and $V^{\pi'}$.

Next, we are interested in bounding the divergence term, $\|d^{\pi'} - d^{\pi}\|_1$. We give the following lemma; to the best of our knowledge, this is a new result.

Lemma 3. *The divergence between discounted future state visitation distributions, $\|d^{\pi'} - d^{\pi}\|_1$, is bounded by an average divergence of the policies π' and π:*

$$\|d^{\pi'} - d^{\pi}\|_1 \leq \frac{2\gamma}{1-\gamma} \underset{s\sim d^{\pi}}{\mathrm{E}} \left[D_{TV}(\pi'||\pi)[s]\right], \tag{25}$$

where $D_{TV}(\pi'||\pi)[s] = (1/2)\sum_a |\pi'(a|s) - \pi(a|s)|$.

Proof. First, using (21), we obtain

$$\begin{aligned} \|d^{\pi'} - d^{\pi}\|_1 &= \gamma\|\bar{G}\Delta d^{\pi}\|_1 \\ &\leq \gamma\|\bar{G}\|_1\|\Delta d^{\pi}\|_1. \end{aligned}$$

$\|\bar{G}\|_1$ is bounded by:

$$\|\bar{G}\|_1 = \|(I - \gamma P_{\pi'})^{-1}\|_1 \leq \sum_{t=0}^{\infty} \gamma^t \|P_{\pi'}\|_1^t = (1-\gamma)^{-1}$$

To conclude the lemma, we bound $\|\Delta d^{\pi}\|_1$.

$$\begin{aligned}
\|\Delta d^{\pi}\|_1 &= \sum_{s'} \left| \sum_{s} \Delta(s'|s) d^{\pi}(s) \right| \\
&\leq \sum_{s,s'} \left|\Delta(s'|s)\right| d^{\pi}(s) \\
&= \sum_{s,s'} \left| \sum_{a} P(s'|s,a) \left(\pi'(a|s) - \pi(a|s)\right) \right| d^{\pi}(s) \\
&\leq \sum_{s,a,s'} P(s'|s,a) \left|\pi'(a|s) - \pi(a|s)\right| d^{\pi}(s) \\
&= \sum_{s,a} \left|\pi'(a|s) - \pi(a|s)\right| d^{\pi}(s) \\
&= 2 \mathop{\mathrm{E}}_{s\sim d^{\pi}} \left[D_{TV}(\pi'||\pi)[s]\right].
\end{aligned}$$

□

The new policy improvement bound follows immediately.

Theorem 1. *For any function $f : S \to \mathbb{R}$ and any policies π' and π, define $\delta_f(s,a,s') \doteq R(s,a,s') + \gamma f(s') - f(s)$,*

$$\epsilon_f^{\pi'} \doteq \max_s \left|\mathrm{E}_{a\sim\pi', s'\sim P}[\delta_f(s,a,s')]\right|,$$

$$L_{\pi,f}(\pi') \doteq \mathop{\mathrm{E}}_{\substack{s\sim d^{\pi} \\ a\sim\pi \\ s'\sim P}} \left[\left(\frac{\pi'(a|s)}{\pi(a|s)} - 1 \right) \delta_f(s,a,s') \right], \text{ and}$$

$$D_{\pi,f}^{\pm}(\pi') \doteq \frac{L_{\pi,f}(\pi')}{1-\gamma} \pm \frac{2\gamma\epsilon_f^{\pi'}}{(1-\gamma)^2} \mathop{\mathrm{E}}_{s\sim d^{\pi}} \left[D_{TV}(\pi'||\pi)[s]\right],$$

where $D_{TV}(\pi'||\pi)[s] = (1/2)\sum_a |\pi'(a|s) - \pi(a|s)|$ is the total variational divergence between action distributions at s. The following bounds hold:

$$D_{\pi,f}^{+}(\pi') \geq J(\pi') - J(\pi) \geq D_{\pi,f}^{-}(\pi'). \tag{4}$$

Furthermore, the bounds are tight (when $\pi' = \pi$, all three expressions are identically zero).

Proof. Begin with the bounds from lemma 2 and bound the divergence $D_{TV}(d^{\pi'}||d^{\pi})$ by lemma 3. □

10.2. Proof of Analytical Solution to LQCLP

Theorem 2 (Optimizing Linear Objective with Linear and Quadratic Constraints)**.** *Consider the problem*

$$\begin{aligned}
p^* = \min_x \; & g^T x \\
\textit{s.t.} \; & b^T x + c \leq 0 \\
& x^T H x \leq \delta,
\end{aligned} \tag{26}$$

where $g, b, x \in \mathbb{R}^n$, $c, \delta \in \mathbb{R}$, $\delta > 0$, $H \in \mathbb{S}^n$, and $H \succ 0$. When there is at least one strictly feasible point, the optimal point x^* satisfies

$$x^* = -\frac{1}{\lambda^*} H^{-1}\left(g + \nu^* b\right),$$

where λ^* and ν^* are defined by

$$\begin{aligned}
\nu^* &= \left(\frac{\lambda^* c - r}{s}\right)_+, \\
\lambda^* &= \arg\max_{\lambda \geq 0} \begin{cases} f_a(\lambda) \doteq \frac{1}{2\lambda}\left(\frac{r^2}{s} - q\right) + \frac{\lambda}{2}\left(\frac{c^2}{s} - \delta\right) - \frac{rc}{s} & \text{if } \lambda c - r > 0 \\ f_b(\lambda) \doteq -\frac{1}{2}\left(\frac{q}{\lambda} + \lambda\delta\right) & \text{otherwise,} \end{cases}
\end{aligned}$$

with $q = g^T H^{-1} g$, $r = g^T H^{-1} b$, and $s = b^T H^{-1} b$.

Furthermore, let $\Lambda_a \doteq \{\lambda | \lambda c - r > 0, \lambda \geq 0\}$, and $\Lambda_b \doteq \{\lambda | \lambda c - r \leq 0, \lambda \geq 0\}$. The value of λ^* satisfies

$$\lambda^* \in \left\{ \lambda_a^* \doteq \text{Proj}\left(\sqrt{\frac{q - r^2/s}{\delta - c^2/s}}, \Lambda_a\right), \lambda_b^* \doteq \text{Proj}\left(\sqrt{\frac{q}{\delta}}, \Lambda_b\right) \right\},$$

with $\lambda^* = \lambda_a^*$ if $f_a(\lambda_a^*) > f_b(\lambda_b^*)$ and $\lambda^* = \lambda_b^*$ otherwise, and Proj(a, S) is the projection of a point x on to a set S. Note: the projection of a point $x \in \mathbb{R}$ onto a convex segment of $\mathbb{R}$, $[a, b]$, has value Proj$(x, [a, b]) = \max(a, \min(b, x))$.

Proof. This is a convex optimization problem. When there is at least one strictly feasible point, strong duality holds by Slater's theorem. We exploit strong duality to solve the problem analytically.

$$\begin{aligned}
p^* &= \min_x \max_{\substack{\lambda \geq 0 \\ \nu \geq 0}} g^T x + \frac{\lambda}{2}\left(x^T H x - \delta\right) + \nu\left(b^T x + c\right) \\
&= \max_{\substack{\lambda \geq 0 \\ \nu \geq 0}} \min_x \frac{\lambda}{2} x^T H x + (g + \nu b)^T x + \left(\nu c - \frac{1}{2}\lambda\delta\right) && \text{Strong duality} \\
&\implies x^* = -\frac{1}{\lambda} H^{-1}(g + \nu b) && \nabla_x \mathcal{L}(x, \lambda, \nu) = 0 \\
&= \max_{\substack{\lambda \geq 0 \\ \nu \geq 0}} -\frac{1}{2\lambda}(g + \nu b)^T H^{-1}(g + \nu b) + \left(\nu c - \frac{1}{2}\lambda\delta\right) && \text{Plug in } x^* \\
&= \max_{\substack{\lambda \geq 0 \\ \nu \geq 0}} -\frac{1}{2\lambda}\left(q + 2\nu r + \nu^2 s\right) + \left(\nu c - \frac{1}{2}\lambda\delta\right) && \text{Notation: } q \doteq g^T H^{-1} g,\ \ r \doteq g^T H^{-1} b,\ \ s \doteq b^T H^{-1} b. \\
&\implies \frac{\partial \mathcal{L}}{\partial \nu} = -\frac{1}{2\lambda}(2r + 2\nu s) + c \\
&\implies \nu = \left(\frac{\lambda c - r}{s}\right)_+ && \text{Optimizing single-variable convex quadratic function over } \mathbb{R}_+ \\
&= \max_{\lambda \geq 0} \begin{cases} \frac{1}{2\lambda}\left(\frac{r^2}{s} - q\right) + \frac{\lambda}{2}\left(\frac{c^2}{s} - \delta\right) - \frac{rc}{s} & \text{if } \lambda \in \Lambda_a \\ -\frac{1}{2}\left(\frac{q}{\lambda} + \lambda\delta\right) & \text{if } \lambda \in \Lambda_b \end{cases} && \text{Notation: } \begin{array}{l} \Lambda_a \doteq \{\lambda | \lambda c - r > 0,\ \ \lambda \geq 0\}, \\ \Lambda_b \doteq \{\lambda | \lambda c - r \leq 0,\ \ \lambda \geq 0\} \end{array}
\end{aligned}$$

Observe that when $c < 0$, $\Lambda_a = [0, r/c)$ and $\Lambda_b = [r/c, \infty)$; when $c > 0$, $\Lambda_a = [r/c, \infty)$ and $\Lambda_b = [0, r/c)$.

Notes on interpreting the coefficients in the dual problem:

- We are guaranteed to have $r^2/s - q \leq 0$ by the Cauchy-Schwarz inequality. Recall that $q = g^T H^{-1} g$, $r = g^T H^{-1} b$, $s = b^T H^{-1} b$. The Cauchy-Scwarz inequality gives:

$$\begin{aligned}
&\|H^{-1/2} b\|_2^2 \|H^{-1/2} g\|_2^2 \geq \left(\left(H^{-1/2} b\right)^T \left(H^{-1/2} g\right)\right)^2 \\
\implies &\left(b^T H^{-1} b\right)\left(g^T H^{-1} g\right) \geq \left(b^T H^{-1} g\right)^2 \\
\therefore\ \ &qs \geq r^2.
\end{aligned}$$

- The coefficient $c^2/s - \delta$ relates to whether or not the plane of the linear constraint intersects the quadratic trust region. An intersection occurs if there exists an x such that $c + b^T x = 0$ with $x^T H x \leq \delta$. To check whether this is the case, we solve
$$x^* = \arg\min_x x^T H x \quad : \quad c + b^T x = 0 \tag{27}$$
and see if $x^{*T} H x^* \leq \delta$. The solution to this optimization problem is $x^* = cH^{-1}b/s$, thus $x^{*T} H x^* = c^2/s$. If $c^2/s - \delta \leq 0$, then the plane intersects the trust region; otherwise, it does not.

If $c^2/s - \delta > 0$ and $c < 0$, then the quadratic trust region lies entirely within the linear constraint-satisfying halfspace, and we can remove the linear constraint without changing the optimization problem. If $c^2/s - \delta > 0$ and $c > 0$, the problem is infeasible (the intersection of the quadratic trust region and linear constraint-satisfying halfspace is empty). Otherwise, we follow the procedure below.

Solving the dual for λ: for any $A > 0$, $B > 0$, the problem

$$\max_{\lambda \geq 0} f(\lambda) \doteq -\frac{1}{2}\left(\frac{A}{\lambda} + B\lambda\right)$$

has optimal point $\lambda^* = \sqrt{A/B}$ and optimal value $f(\lambda^*) = -\sqrt{AB}$.

We can use this solution form to obtain the optimal point on each segment of the piecewise continuous dual function for λ:

objective	optimal point (before projection)	optimal point (after projection)
$f_a(\lambda) \doteq \frac{1}{2\lambda}\left(\frac{r^2}{s} - q\right) + \frac{\lambda}{2}\left(\frac{c^2}{s} - \delta\right) - \frac{rc}{s}$	$\lambda_a \doteq \sqrt{\frac{q - r^2/s}{\delta - c^2/s}}$	$\lambda_a^* = \text{Proj}(\lambda_a, \Lambda_a)$
$f_b(\lambda) \doteq -\frac{1}{2}\left(\frac{q}{\lambda} + \lambda\delta\right)$	$\lambda_b \doteq \sqrt{\frac{q}{\delta}}$	$\lambda_b^* = \text{Proj}(\lambda_b, \Lambda_b)$

The optimization is completed by comparing $f_a(\lambda_a^*)$ and $f_b(\lambda_b^*)$:

$$\lambda^* = \begin{cases} \lambda_a^* & f_a(\lambda_a^*) \geq f_b(\lambda_b^*) \\ \lambda_b^* & \text{otherwise.} \end{cases}$$

□

10.3. Experimental Parameters

10.3.1. Environments

In the Circle environments, the reward and cost functions are

$$R(s) = \frac{v^T[-y, x]}{1 + |\|[x, y]\|_2 - d|},$$
$$C(s) = \mathbf{1}\left[|x| > x_{lim}\right],$$

where x, y are the coordinates in the plane, v is the velocity, and d, x_{lim} are environmental parameters. We set these parameters to be

	Point-mass	Ant	Humanoid
d	15	10	10
x_{lim}	2.5	3	2.5

In Point-Gather, the agent receives a reward of $+10$ for collecting an apple, and a cost of 1 for collecting a bomb. Two apples and eight bombs spawn on the map at the start of each episode. In Ant-Gather, the reward and cost structure was the same, except that the agent also receives a reward of -10 for falling over (which results in the episode ending). Eight apples and eight bombs spawn on the map at the start of each episode.

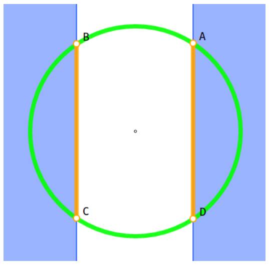

Figure 5. In the Circle task, reward is maximized by moving along the green circle. The agent is not allowed to enter the blue regions, so its optimal constrained path follows the line segments AD and BC.

10.3.2. Algorithm Parameters

In all experiments, we use Gaussian policies with mean vectors given as the outputs of neural networks, and with variances that are separate learnable parameters. The policy networks for all experiments have two hidden layers of sizes $(64, 32)$ with $\tanh$ activation functions.

We use GAE-λ (Schulman et al., 2016) to estimate the advantages and constraint advantages, with neural network value functions. The value functions have the same architecture and activation functions as the policy networks. We found that having different λ^{GAE} values for the regular advantages and the constraint advantages worked best. We denote the λ^{GAE} used for the constraint advantages as λ_C^{GAE}.

For the failure prediction networks $P_\phi(s \to U)$, we use neural networks with a single hidden layer of size (32), with output of one sigmoid unit. At each iteration, the failure prediction network is updated by some number of gradient descent steps using the Adam update rule to minimize the prediction error. To reiterate, the failure prediction network is a model for the probability that the agent will, at some point in the next T time steps, enter an unsafe state. The cost bonus was weighted by a coefficient α, which was 1 in all experiments except for Ant-Gather, where it was 0.01. Because of the short time horizon, no cost bonus was used for Point-Gather.

For all experiments, we used a discount factor of $\gamma = 0.995$, a GAE-λ for estimating the regular advantages of $\lambda^{GAE} = 0.95$, and a KL-divergence step size of $\delta_{KL} = 0.01$.

Experiment-specific parameters are as follows:

Parameter	Point-Circle	Ant-Circle	Humanoid-Circle	Point-Gather	Ant-Gather
Batch size	50,000	100,000	50,000	50,000	100,000
Rollout length	50-65	500	1000	15	500
Maximum constraint value d	5	10	10	0.1	0.2
Failure prediction horizon T	5	20	20	(N/A)	20
Failure predictor SGD steps per itr	25	25	25	(N/A)	10
Predictor coeff α	1	1	1	(N/A)	0.01
λ_C^{GAE}	1	0.5	0.5	1	0.5

Note that these same parameters were used for all algorithms.

We found that the Point environment was agnostic to λ_C^{GAE}, but for the higher-dimensional environments, it was necessary to set λ_C^{GAE} to a value < 1. Failing to discount the constraint advantages led to substantial overestimates of the constraint gradient magnitude, which led the algorithm to take unsafe steps. The choice $\lambda_C^{GAE} = 0.5$ was obtained by a hyperparameter search in $\{0.5, 0.92, 1\}$, but 0.92 worked nearly as well.

10.3.3. Primal-Dual Optimization Implementation

Our primal-dual implementation is intended to be as close as possible to our CPO implementation. The key difference is that the dual variables for the constraints are stateful, learnable parameters, unlike in CPO where they are solved from scratch at each update.

The update equations for our PDO implementation are

$$\begin{aligned}\theta_{k+1} &= \theta_k + s^j \sqrt{\frac{2\delta}{(g - \nu_k b)^T H^{-1} (g - \nu_k b)}} H^{-1} \left(g - \nu_k b\right) \\ \nu_{k+1} &= \left(\nu_k + \alpha \left(J_C(\pi_k) - d\right)\right)_+ ,\end{aligned}$$

where s^j is from the backtracking line search ($s \in (0,1)$ and $j \in \{0, 1, ..., J\}$, where J is the backtrack budget; this is the same line search as is used in CPO and TRPO), and α is a learning rate for the dual parameters. α is an important hyperparameter of the algorithm: if it is set to be too small, the dual variable won't update quickly enough to meaningfully enforce the constraint; if it is too high, the algorithm will overcorrect in response to constraint violations and behave too conservatively. We experimented with a relaxed learning rate, $\alpha = 0.001$, and an aggressive learning rate, $\alpha = 0.01$. The aggressive learning rate performed better in our experiments, so all of our reported results are for $\alpha = 0.01$.

Selecting the correct learning rate can be challenging; the need to do this is obviated by CPO.