Appendix

We now provide the proofs of our lemmas.

Impact on optimization speed

To prove Lemma 1, we first need to prove an additional lemma.

Lemma 4 (Set of Cramér extensions)**.** *The set $\mathcal{C}$ of symmetric matrices M such that $(\mathbf{p}-\mathbf{q})^\top M(\mathbf{p}-\mathbf{q}) = (\mathbf{p}-\mathbf{q})^\top CC^\top(\mathbf{p}-\mathbf{q})$ for all normalized distributions $\mathbf{p}$ and $\mathbf{q}$ is the set*

$$\mathcal{C} = \left\{CC^\top + ae^\top + ea^\top | a \in \mathbb{R}^k\right\} \ .$$

Proof. Let $M = CC^\top + ae^\top + ea^\top$. Since $\mathbf{p}$ and $\mathbf{q}$ are normalized, we have $(\mathbf{p}-\mathbf{q})^\top e = 0$. Hence,

$$\begin{aligned}&(\mathbf{p}-\mathbf{q})^\top M(\mathbf{p}-\mathbf{q}) = \\ &\quad (\mathbf{p}-\mathbf{q})^\top CC^\top(\mathbf{p}-\mathbf{q}) \\ &\quad + (\mathbf{p}-\mathbf{q})^\top ae^\top(\mathbf{p}-\mathbf{q}) + (\mathbf{p}-\mathbf{q})^\top ea^\top(\mathbf{p}-\mathbf{q}) \\ &= (\mathbf{p}-\mathbf{q})^\top CC^\top(\mathbf{p}-\mathbf{q}) \ .\end{aligned}$$

Conversely, let a symmetric matrix M be such that $(\mathbf{p}-\mathbf{q})^\top M(\mathbf{p}-\mathbf{q}) = (\mathbf{p}-\mathbf{q})^\top CC^\top(\mathbf{p}-\mathbf{q})$ for all normalized $\mathbf{p}$ and $\mathbf{q}$. Then $(\mathbf{p}-\mathbf{q})^\top(M - CC^\top)(\mathbf{p}-\mathbf{q}) = 0$. For this to be true, $(M - CC^\top)(\mathbf{p}-\mathbf{q})$ must be colinear to e. Thus, denoting $M - CC^\top = ea^\top + N$ where e is not in the span of N, we must have $N(\mathbf{p}-\mathbf{q}) = 0$ for all normalized $\mathbf{p}$ and $\mathbf{q}$, i.e. $N = be^\top$. The symmetry constraint leads to $a = b$. This concludes the proof. □

Lemma 1 (Condition number)**.** *Let $\mathcal{C}$ be the set of symmetric matrices M for which $(\mathbf{p}-\mathbf{q})^\top M(\mathbf{p}-\mathbf{q}) = (\mathbf{p}-\mathbf{q})^\top CC^\top(\mathbf{p}-\mathbf{q})$ for all proper distributions $\mathbf{p}$ and $\mathbf{q}$. Let $\kappa_{\min}(\mathcal{C})$ the lowest condition number attained by matrices M in $\mathcal{C}$. Then all the matrices of the form C_λ with $\lambda \in [\lambda_{k-1}(C_0), \lambda_1(C_0)]$, where $\lambda_{k-1}(C_0)$ and $\lambda_1(C_0)$ are the second smallest and largest eigenvalues of C_0, respectively, have condition number $\kappa_{\min}(\mathcal{C})$.*

Proof. Let v_L be the vector associated with the maximum eigenvalue L of C_0 and a be an arbitrary vector. Because C_0 is a symmetric matrix whose only zero eigenvalue corresponds to e, its eigenvectors are orthogonal to e and in particular $v_L^\top e = 0$. Thus, we have

$$\begin{aligned}L &= v_L^\top C_0 v_L \\ &= v_L^\top \Pi_{e^\perp} CC^\top \Pi_{e^\perp} v_L \\ &= v_L^\top CC^\top v_L \\ &= v_L^\top CC^\top v_L + v_L^\top ae^\top v_L + v_L^\top ea^\top v_L\end{aligned}$$

for any vector a since $e^\top v_L = 0$. Denoting $R_a = CC^\top + ae^\top + ea^\top$, we get

$$\begin{aligned}L &= v_L^\top R_a v_L \\ &\leq \max_v \frac{v^\top R_a v}{\|v\|^2} \ ,\end{aligned}$$

which is the largest eigenvalue of R_a. Since this is true for every a, C_0 has the lowest top eigenvalue from all the matrices in $\mathcal{C}$.

Similarly, let us denote v_μ be the vector associated with the second-smallest eigenvalue μ [1]. As e is the eigenvector associated with the eigenvalue 0, we have that $v_\mu^\top e = 0$ and

$$\begin{aligned}\mu &= v_\mu^\top C_0 v_\mu \\ &= v_\mu^\top CC^\top v_\mu \\ &= v_\mu^\top CC^\top v_\mu + v_\mu^\top ae^\top v_\mu + v_\mu^\top ea^\top v_\mu \\ &= v_\mu^\top R_a v_\mu \\ &\geq \min_v \frac{v^\top R_a v}{\|v\|^2} \ .\end{aligned}$$

Thus, for all a, the second smallest eigenvalue of C_0 is larger than the smallest eigenvalue of R_a.

This means that, for C_λ to have the smallest condition number of all the matrices in $\mathcal{C}$, it is sufficient to require that the eigenvalue associated with e be between μ and L, i.e. that $\mu \leq \lambda \leq L$. This concludes the proof. □

Preservation of the expectation

To prove Lemma 2, we will need the following proposition:

Proposition 1. *Let $\mathbf{z}$ be defined as in Section 3, i.e. $\mathbf{z}$ is the vector of evenly spaced returns between $-\frac{k-1}{2}$ and $\frac{k-1}{2}$ with mean 0. Let $b = [-1, 0, 0, \ldots, 0, 0, 1]^\top$. Then $C_\lambda^{-1}\mathbf{z} = b$ for all values of $\lambda > 0$.*

Proof. We will prove that $C_\lambda b = \mathbf{z}$ for all values of λ. First, we note that $e^\top b = 0$ and $C_\lambda b = \Pi_{e^\perp} CC^\top b$.

Since $C_{ij} = 1_{i\geq j}$, we have, denoting $c = C^\top b$,

$$\begin{aligned}c_j &= \sum_i C_{ij} b_i \\ &= \begin{cases} 0 & \text{if j = 1} \\ 1 & \text{otherwise} \end{cases} \ .\end{aligned}$$

Multiplying by C to get $d = Cc$, we get

$$\begin{aligned}d_i &= \sum_j C_{ij}(C^\top b)_j \\ &= i - 1 \ .\end{aligned}$$

[1] The smallest being 0.

We now need to compute $r = \Pi_{e^\perp} d$. Since $e^\top d = \frac{(k-1)\sqrt{k}}{2}$, we have

$$\begin{aligned} r_i &= d_i - V \\ &= i - 1 - \frac{k-1}{2} \\ &= \frac{2i - 1 - k}{2} \\ &= \mathbf{z}_i \ . \end{aligned}$$

This concludes the proof. □

Lemma 2 (Expectation preserving)**.** *Let* $\mathbf{p}$ *be an arbitrary distribution over a discrete support. Let* $\Pi_{A,b}(\mathbf{p})$ *the projection of* $\mathbf{p}$ *onto the linear subset* $\mathcal{S}_{A,b} = \{\mathbf{q} | A\mathbf{q} = b\}$*. Then, if the first and the last columns of* A *are equal, i.e.* $A_1 = A_k$*, then* $\mathbf{p}$ *and* $\Pi_{A,b}(\mathbf{p})$ *have the same expectation.*

Proof. By definition,

$$\Pi_{A,b}(\mathbf{p}) = \begin{array}{ll} \arg\min_{\mathbf{q}} & (\mathbf{p} - \mathbf{q})^\top C_\lambda (\mathbf{p} - \mathbf{q}) \\ \text{subject to} & A\mathbf{q} = b \ . \end{array}$$

Writing ν the Lagrange multipliers, this is a quadratic program whose solution is given by

$$\left[\begin{array}{c} \Pi_{A,b}(\mathbf{p}) \\ \nu \end{array} \right] = \left[\begin{array}{cc} C_\lambda & A^\top \\ A & 0 \end{array} \right]^{-1} \left[\begin{array}{c} C_\lambda \mathbf{p} \\ b \end{array} \right] \ .$$

Inverting the block diagonal matrix yields

$$\left[\begin{array}{cc} C_\lambda & A^\top \\ A & 0 \end{array} \right]^{-1} = \left[\begin{array}{cc} M_{11} & M_{12} \\ M_{21} & M_{22} \end{array} \right]$$

with

$$\begin{aligned} M_{11} &= C_\lambda^{-1} - C_\lambda^{-1} A^\top (A C_\lambda^{-1} A^\top)^{-1} A C_\lambda^{-1} \\ M_{12} &= C_\lambda^{-1} A^\top (A C_\lambda^{-1} A^\top)^{-1} \\ M_{21} &= (A C_\lambda^{-1} A^\top)^{-1} A C_\lambda^{-1} \\ M_{21} &= -(A C_\lambda^{-1} A^\top)^{-1} \ . \end{aligned}$$

Hence,

$$\begin{aligned} \Pi_{A,b}(\mathbf{p}) &= M_{11} C_\lambda \mathbf{p} + M_{12} b \\ &= \mathbf{p} - C_\lambda^{-1} A^\top s \end{aligned}$$

for some s. Thus, the expected Q-value with respect to the projected distribution is equal to

$$\mathbf{z}^\top \Pi_{A,b}(\mathbf{p}) = \mathbf{z}^\top \mathbf{p} - \mathbf{z}^\top C_\lambda^{-1} A^\top s$$

and the two expectations will be equal if $\mathbf{z}^\top C_\lambda^{-1} A^\top s = 0$. Using Proposition 1, we know that $C_\lambda^{-1} \mathbf{z} = b$. Thus, if it sufficient to have $Ab = 0$ for the two expectations to match. Since only the first and the last components of b are nonzeros and they are opposite of each other, we have $Ab = 0 \Leftrightarrow A_1 = A_k$ when denoting A_j the j-th column of A. This concludes the proof. □

Convergence to a fixed point

This result requires additional definitions. A value distribution $\mathbf{P}$ maps states $x \in \mathcal{X}$ to distributions on $\mathbb{R}$; we extend this to vectors defined by a linear combination of features:

$$\mathbf{P}(x) = \Theta^\top \phi(x) \ ,$$

where $\phi(x) \in \mathbb{R}^m$ is the feature vector at state x and $\Theta \in \mathbb{R}^{m \times k}$ is the parameter matrix we try to estimate.

Concatening all feature vectors into a feature matrix $\Phi \in \mathbb{R}^{n \times m}$, our linear approximation is $\mathbf{P}_\Theta := \Phi\Theta \in \mathbb{R}^{n \times k}$. We assume that the vector $\mathbf{P}_\Theta(x) \in \mathbb{R}^k$ approximates a distribution over the support $\mathbf{z} := \{z_1, z_2, \ldots, z_k\}$, but it may have negative components and is not necessarily normalized.

We are given a distribution ξ on $\mathcal{X}$ and we shall use a Cramér distance between distributions over $\mathbf{z}$:

$$l_\lambda^2(\mathbf{p}, \mathbf{q}) := \|\mathbf{p} - \mathbf{q}\|_{C_\lambda}^2 \ .$$

We transform the matrix C_λ into an operator over continuous distributions, where with some abuse of notation we view $\mathbf{p}$ as a distribution over a finite set of Diracs: $\mathbf{p}(y) = \sum_i p_i \delta_{z_i = y}$. Then

$$\begin{aligned} \Pi_{e^\perp} \mathbf{p}(x) &= \mathbf{p}(x) - \int_{y=z_1}^{z_k} \mathbf{p}(y) \, dy \\ \Pi_{e^\perp} \mathbf{q}(x) &= \mathbf{q}(x) - \int_{y=z_1}^{z_k} \mathbf{q}(y) \, dy \\ l_\lambda^2(\mathbf{p}, \mathbf{q}) &= \int_{x=z_1}^{z_k} \left(\int_{y=z_1}^{x} \left[\Pi_{e^\perp} \mathbf{p}(y) - \Pi_{e^\perp} \mathbf{q}(y) \right] \, dx \right)^2 dy \\ &\quad + \lambda \left(\int_{y=z_1}^{z_k} \left[\mathbf{p}(y) - \mathbf{q}(y) \right] \, dx \right)^2 \ . \end{aligned} \tag{10}$$

The first term on the right-hand side of Eq. (10) penalizes the difference in cdf of $\mathbf{p}$ and $\mathbf{q}$ while the second term penalizes the difference in mass. When applied to two distributions $\mathbf{p}$ and $\mathbf{q}$ over $\mathbf{z}$, this is equivalent to $(\mathbf{p} - \mathbf{q})^\top C_\lambda (\mathbf{p} - \mathbf{q})$. We define the weighted Cramér distance over value distributions by

$$l_{\xi,\lambda}^2(\mathbf{P}, \mathbf{Q}) := \sum_{x \in \mathcal{X}} \xi(x) l_\lambda^2(\mathbf{P}(x), \mathbf{Q}(x)).$$

In what follows we identify three spaces of distributions or distribution-like objects. First, $\mathbb{P}$ is the space of distributions with support the interval $[z_1, z_k)$. $\mathcal{D}$ is the space of distributions over $\mathbf{z}$. Finally, $\mathcal{P}$ is the vector space spanned by the features $\Phi \in \mathbb{R}^{n \times m}$, that is: $\mathcal{P} = \{\Phi\Theta : \Theta \in \mathbb{R}^{m \times k}\}$.

While our value distribution will only output distributions over the support $\mathbf{z}$, the distributional Bellman operator $\mathcal{T}^\pi$ transforms distributions over $\mathbf{z}$ into distributions from $\mathbb{P}$. We thus need to consider the projection

$\Pi_{\lambda,\mathcal{P}}$ which projects $\mathbb{P}$ onto $\mathcal{D}$:

$$\Pi_{\lambda,\mathcal{D}}\mathbf{p} = \arg\min_{\mathbf{q}\in\mathcal{D}} l_\lambda^2(\mathbf{p},\mathbf{q}) \ .$$

Lemma 3 from Rowland et al. (2018) states that, for any distribution $\mathbf{p} \in \mathcal{D}$, we have

$$l_\lambda^2(\mathbf{p},\mathbf{q}) = l_\lambda^2(\mathbf{p},\Pi_{\lambda,\mathcal{D}}\mathbf{p}) + l_\lambda^2(\Pi_{\lambda,\mathcal{D}}\mathbf{p},\mathbf{q}). \tag{11}$$

We now move from the projection of distributions to the projection of value distributions. We define a projection in $l_{\xi,\lambda}^2$ of a value distribution $\mathbf{Q}$ onto the subspace $\mathcal{V}$ by

Definition 1. *The ξ-weighted projection onto $\mathcal{V}$ is*

$$\mathbf{\Pi}_{\xi,\lambda,\mathcal{V}}\mathbf{P} := \arg\min_{\mathbf{Q}\in\mathcal{V}} l_{\xi,\lambda}^2(\mathbf{P},\mathbf{Q}) \ ,$$

where both the projection and the distance are in bold to distinguish them from projection and distances in distribution space.

In particular, two projections are of interest. First, we consider the set of all value distributions from $\mathcal{X}$ to distributions supported by $\mathbf{z}$. The projection onto this set is

$$[\mathbf{\Pi}_{\xi,\lambda,\mathcal{D}}\mathbf{P}](x) = \Pi_{\lambda,\mathcal{D}}\mathbf{P}(x) \ ,$$

We are also interested in the ξ-weighted projection onto Φ, the set of linear value distributions:

$$\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P} := \arg\min_{\Phi\Theta,\Theta\in\mathbb{R}^{m\times k}} l_{\xi,\lambda}^2(\mathbf{P},\Theta\Phi),$$

The projection $\mathbf{\Pi}_{\xi,\lambda,\Phi}$ of the true value distribution $\mathbf{Q}$ gives us the closest linear value distribution according to the Cramér distance defined by C_λ.

Lemma 5 (Projection onto Φ)**.** *Let $\mathbf{P}$ be an arbitrary value distribution supported on $\mathbb{P}$. The ξ-weighted projection of $\mathbf{P}$ onto Φ, $\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}$, is equal to the ξ-weighted projection of $\mathbf{\Pi}_{\xi,\lambda,\mathcal{D}}\mathbf{P}$.*

The above lemma will let us restrict our attention to distributions on $\mathbf{z}$, that is $\mathbf{P} \in \mathcal{D}$.

Proof. Fix $\mathbf{Q} := \Phi\Theta$. By definition, the support of $\mathbf{Q}$ is $\mathcal{P}$. Now

$$\begin{aligned} l_{\xi,\lambda}^2(\mathbf{P},\mathbf{Q}) &= \sum_{x\in\mathcal{X}} \xi(x) l_\lambda^2(\mathbf{P}(x),\mathbf{Q}(x)) \\ &= \sum_{x\in\mathcal{X}} \xi(x) l_\lambda^2(\mathbf{Q}(x),\Pi_{\lambda,\mathcal{D}}\mathbf{P}(x)) \\ &\quad + \sum_{x\in\mathcal{X}} \xi(x) l_\lambda^2(\Pi_{\lambda,\mathcal{D}}\mathbf{P}(x),\mathbf{P}(x)) \\ &= l_{\xi,\lambda}^2(\mathbf{Q},\Pi_{\lambda,\mathcal{D}}\mathbf{P}) + \\ &\qquad \sum_{x\in\mathcal{X}} \xi(x) l_\lambda^2(\Pi_{\lambda,\mathcal{D}}\mathbf{P}(x),\mathbf{P}(x)), \end{aligned}$$

using Eq. 11. From the above we deduce that the matrix Θ which minimizes $l_{\xi,\lambda}^2(\Phi\Theta,\mathbf{P})$ is also the minimizer of $l_{\xi,\lambda}^2(\Phi\Theta,\Pi_{\lambda,\mathcal{D}}\mathbf{P})$. □

Lemma 6 ($\mathbf{\Pi}_{\xi,\lambda,\Phi}$ is a non-expansion)**.** $\mathbf{\Pi}_{\xi,\lambda,\Phi}$ *is a non-expansion in $l_{\xi,\lambda}^2$, i.e. for every pair $(\mathbf{P},\mathbf{Q})$ of value distributions, we have*

$$l_{\xi,\lambda}^2(\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P},\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{Q}) \le l_{\xi,\lambda}^2(\mathbf{P},\mathbf{Q}) \ .$$

Proof. We can view $l_{\xi,\lambda}^2$ as a weighted L_2 norm over vectors in $\mathbb{R}^{n\times k}$, with $\mathbf{\Pi}_{\xi,\lambda,\Phi}$ the corresponding projection onto the affine subspace spanned by Φ. The result is standard from these observations. □

Recall that the loss l_λ^2 between vectors is defined through the matrix $C_\lambda = \Pi_{e^\perp}CC^\top\Pi_{e^\perp} + \lambda ee^\top$: $l_\lambda^2(\mathbf{p},\mathbf{q}) = (\mathbf{p}-\mathbf{q})^\top C_\lambda(\mathbf{p}-\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_{C_\lambda}^2$. To prove Theorem 1 we will consider two separate components of that loss: along e and along the subspace orthogonal to e. That is, let us write

$$A := \Pi_{e^\perp}C,$$

such that

$$l_\lambda^2(\mathbf{p},\mathbf{q}) = \|\mathbf{p}-\mathbf{q}\|_{AA^\top}^2 + \lambda\,\|\mathbf{p}-\mathbf{q}\|_{ee^\top}^2 \ .$$

We extend this notation to a ξ-weighted norm over value distributions. For a matrix $B \in \mathbb{R}^{k\times k}$ and $\Delta \in \mathbb{R}^{n\times k}$ write

$$\|\Delta\|_{\xi,B}^2 = \sum_{x\in\mathcal{X}} \xi(x)\,\|\Delta(x)\|_B^2 \ ,$$

where we associate each state $x \in \mathcal{X}$ with an integer in $\{1,\ldots,n\}$. Then:

$$l_{\xi,\lambda}^2(\mathbf{P},\mathbf{Q}) = \|\mathbf{P}-\mathbf{Q}\|_{\xi,AA^\top}^2 + \lambda\,\|\mathbf{P}-\mathbf{Q}\|_{\xi,ee^\top}^2 \ .$$

Lemma 3. *Let ξ be the stationary distribution induced by the policy π. Write $\mathcal{T}^{\pi\prime} := \Pi_{\lambda,\mathcal{D}}\mathcal{T}^\pi$ to mean the distributional Bellman operator followed by a projection onto the support $\mathbf{z} = z_1,\ldots,z_k$. For a matrix $B \in \mathbb{R}^{k\times k}$ and $\Delta \in \mathbb{R}^{n\times k}$, write*

$$\|\Delta\|_{\xi,B}^2 = \sum_{x\in\mathcal{X}} \xi(x)\,\|\Delta(x)\|_B^2 \ .$$

Then for any two value distributions $\mathbf{P},\mathbf{Q} \in \mathbb{R}^{n\times k}$,

$$\begin{aligned} \left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|_{\xi,AA^\top}^2 &\le \gamma\,\|\mathbf{P}-\mathbf{Q}\|_{\xi,AA^\top}^2 \\ \left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|_{\xi,ee^\top}^2 &\le \|\mathbf{P}-\mathbf{Q}\|_{\xi,ee^\top}^2 \ . \end{aligned}$$

where $A := \Pi_{e^\perp}C$.

Lemma 3 states that the distributional Bellman operator, applied over distributions in $\mathbb{R}^{n\times k}$, contracts all dimensions orthogonal to e by a factor $\gamma^{1/2}$ but is only a nonexpansion along e.

Proof. Let $\mathbf{P}, \mathbf{Q}$ be two value distributions. To keep the notation light, without loss of generality let $\lambda = 1$. We begin with the term in $ee^\top$:

$$\begin{aligned}
&\left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|^2_{\xi,ee^\top} \\
&\quad = \sum_{x\in\mathcal{X}} \xi(x) \left\|\mathcal{T}^{\pi\prime}\mathbf{P}(x) - \mathcal{T}^{\pi\prime}\mathbf{Q}(x)\right\|^2_{ee^\top} \\
&\quad = \sum_{x\in\mathcal{X}} \xi(x) \left\|e^\top\mathcal{T}^{\pi\prime}\mathbf{P}(x) - e^\top\mathcal{T}^{\pi\prime}\mathbf{Q}(x)\right\|^2 .
\end{aligned}$$

The term $e^\top\mathcal{T}^{\pi\prime}\mathbf{P}(x)$ measures the total mass at x (up to a multiplicative constant $\sqrt{k}$), after applying the distributional Bellman operator $\mathcal{T}^\pi$ and projecting onto the finite support. $\mathcal{T}^\pi\mathbf{P}(x)$ consists of a mixture of next-state distributions, shifted by the reward $r(x)$ and scaled by the discount factor γ. However, neither of these two operations affects the mass of the distributions. Furthermore, the Cramér projection onto the support also preserves mass (Rowland et al., 2018). Hence

$$e^\top\mathcal{T}^{\pi\prime}\mathbf{P}(x) = \sum_{x'\in\mathcal{X}} \mathrm{Pr}_\pi(x' \,|\, x) e^\top\mathbf{P}(x').$$

And therefore

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} \xi(x) \left\|\mathcal{T}^{\pi\prime}\mathbf{P}(x) - \mathcal{T}^{\pi\prime}\mathbf{Q}(x)\right\|^2_{ee^\top} \\
&= \sum_{x\in\mathcal{X}} \xi(x) \Bigg(\sum_{x'\in\mathcal{X}} \mathrm{Pr}_\pi(x' \,|\, x) e^\top\mathbf{P}(x') - \\
&\qquad\qquad \mathrm{Pr}_\pi(x' \,|\, x) e^\top\mathbf{Q}(x')\Big)^2 \\
&= \sum_{x\in\mathcal{X}} \xi(x) \left(\sum_{x'\in\mathcal{X}} \mathrm{Pr}_\pi(x' \,|\, x) e^\top(\mathbf{P}(x') - \mathbf{Q}(x'))\right)^2 .
\end{aligned}$$

Now, by Jensen's inequality and the fact that $\xi(x') = \sum_x \xi(x)\,\mathrm{Pr}_\pi(x' \,|\, x)$,

$$\begin{aligned}
&\sum_{x\in\mathcal{X}} \xi(x) \big(\sum_{x'\in\mathcal{X}} \mathrm{Pr}_\pi(x' \,|\, x) e^\top(\mathbf{P}(x') - \mathbf{Q}(x'))\big)^2 \\
&\quad \le \sum_{x\in\mathcal{X}} \xi(x) \sum_{x'\in\mathcal{X}} \mathrm{Pr}_\pi(x' \,|\, x) (e^\top(\mathbf{P}(x') - \mathbf{Q}(x')))^2 \\
&\quad = \sum_{x'\in\mathcal{X}} \xi(x')\,\mathrm{Pr}_\pi(x' \,|\, x) (e^\top(\mathbf{P}(x') - \mathbf{Q}(x')))^2 \\
&\quad = \left\|\mathbf{P} - \mathbf{Q}\right\|^2_{\xi,ee^\top} .
\end{aligned}$$

This proves the second statement. For the first, notice that we can add any constant vector $\alpha(x)e$ to the distribution at each state, without changing the $AA^\top$-distance between them:

$$\left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|^2_{\xi,AA^\top} = \left\|\mathcal{T}^{\pi\prime}(\mathbf{P} + \alpha e) - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|^2_{\xi,AA^\top} .$$

In particular, we can choose αe so that the two value distributions have equal mass at all states (and in fact, sum to 1 at all states, by also changing $\mathbf{Q}$). In turn we can modify results by Bellemare et al. (2017b) and Rowland et al. (2018) showing that the distributional Bellman operator, projected onto a finite support or not, is a $\gamma^{1/2}$ contraction in Cramér metric, extending it as above to deal with the ξ-weighted norm rather than the maximal norm. We conclude that

$$\left\|\mathcal{T}^{\pi\prime}\mathbf{P} - \mathcal{T}^{\pi\prime}\mathbf{Q}\right\|^2_{\xi,AA^\top} \le \gamma \left\|\mathbf{P} - \mathbf{Q}\right\|^2_{\xi,AA^\top} . \qquad \square$$

Theorem 1 (Convergence of the projected distributional Bellman process)**.** *Let ξ be the stationary distribution induced by the policy π. The process*

$$\mathbf{P}_0 := \Phi\Theta_0 \quad , \quad \mathbf{P}_{k+1} := \hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{P}_k .$$

converges to a set S such that for any two $\mathbf{P}, \mathbf{P}' \in S$, there is a $\mathcal{X}$-indexed vector of constants α such that

$$\mathbf{P}(x) = \mathbf{P}'(x) + \alpha(x)e.$$

If $\lambda > 0$, S consists of a single point $\tilde{\mathbf{P}}$ which is the fixed point of the process. Furthermore, we can bound the error of this fixed point with respect to the true value distribution $\mathbf{P}^\pi$:

$$\begin{aligned}
l^2_{\xi,\lambda}(\tilde{\mathbf{P}}, \mathbf{P}^\pi) \le \frac{1}{1-\gamma} & l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi, \mathbf{P}^\pi) \\
& - \frac{\gamma\lambda}{1-\gamma} \left\|\tilde{\mathbf{P}} - \mathbf{P}^\pi\right\|^2_{\xi,ee^\top},
\end{aligned}$$

where the second term measures the difference in mass between $\tilde{\mathbf{P}}$ and $\mathbf{P}^\pi$.

Proof (Sketch). To prove the theorem, we cannot make direct use of the usual techniques e.g. from Tsitsiklis & Van Roy (1997). First, the operator $\hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}$ is not a projection operator when $\lambda > 0$, because of the normalization term $\lambda(\mathbf{q}^\top e - 1)^2$ (Equation 9). Second, the Bellman operator is not a contraction when applied to distributions with varying mass.

Let us consider two process $\mathbf{P}_{k+1} = \hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{P}_k$ and $\mathbf{Q}_k = \hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi Q_k$, possibly with different initial conditions. We make use of the following fact:

$$\hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{P} = \mathbf{\Pi}_{\xi,\lambda,\Phi}\tilde{\mathcal{T}}^\pi\mathbf{P},$$

where $\tilde{\mathcal{T}}^\pi\mathbf{P} = \Pi_{e^\perp}\mathcal{T}^\pi\mathbf{P} + \frac{e}{\sqrt{k}}$ is a modification of the distributional Bellman operator which "resets" the mass of the resulting distribution to 1 by adding the appropriate constant vector (recall $e = [1/\sqrt{k}, \ldots, 1/\sqrt{k}]^\top$). We use this fact to measure how the two processes evolve under the norm $\|\cdot\|_{\xi,C_\lambda}$:

$$\begin{aligned}
&\|\mathbf{P}_{k+1} - \mathbf{Q}_{k+1}\|^2_{\xi,C_\lambda} = \\
&\quad = \left\|\hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{P}_k - \hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{Q}_k\right\|^2_{\xi,C_\lambda} \\
&\quad = \left\|\mathbf{\Pi}_{\xi,\lambda,\Phi}\tilde{\mathcal{T}}^\pi\mathbf{P}_k - \mathbf{\Pi}_{\xi,\lambda,\Phi}\tilde{\mathcal{T}}^\pi\mathbf{Q}_k\right\|^2_{\xi,C_\lambda} \\
&\quad \le \left\|\tilde{\mathcal{T}}^\pi\mathbf{P}_k - \tilde{\mathcal{T}}^\pi\mathbf{Q}_k\right\|^2_{\xi,C_\lambda} \\
&\quad = \left\|\tilde{\mathcal{T}}^\pi\mathbf{P}_k - \tilde{\mathcal{T}}^\pi\mathbf{Q}_k\right\|^2_{\xi,AA^\top} + \left\|\tilde{\mathcal{T}}^\pi\mathbf{P}_k - \tilde{\mathcal{T}}^\pi\mathbf{Q}_k\right\|^2_{\xi,ee^\top} \\
&\quad = \|\mathcal{T}^\pi\mathbf{P}_k - \mathcal{T}^\pi\mathbf{Q}_k\|^2_{\xi,AA^\top} + \|\mathcal{T}^\pi\mathbf{P}_k - \mathcal{T}^\pi\mathbf{Q}_k\|^2_{\xi,ee^\top},
\end{aligned}$$

where the last line follows from the fact that the addition of the constant $e/\sqrt{k}$ does not impact either term. Furthermore,

$$\begin{aligned}
\left\|\tilde{\mathcal{T}}^\pi\mathbf{P}_k - \tilde{\mathcal{T}}^\pi\mathbf{Q}_k\right\|^2_{\xi,ee^\top} &= \|\Pi_{e^\perp}\mathbf{P}_k - \Pi_{e^\perp}\mathbf{Q}_k\|^2_{\xi,ee^\top} \\
&= 0.
\end{aligned}$$

It follows from Lemma 3 that

$$\begin{aligned}
\|\mathbf{P}_{k+1} - \mathbf{Q}_{k+1}\|^2_{\xi,C_\lambda} &\le \left\|\tilde{\mathcal{T}}^\pi\mathbf{P}_k - \tilde{\mathcal{T}}^\pi\mathbf{Q}_k\right\|^2_{\xi,AA^\top} \\
&\le \gamma\,\|\mathbf{P}_k - \mathbf{Q}_k\|^2_{\xi,AA^\top} \\
&\le \gamma\,\|\mathbf{P}_k - \mathbf{Q}_k\|^2_{\xi,C_\lambda}.
\end{aligned}$$

Now if $\lambda > 0$, the norm $\|\cdot\|_{\xi,C_\lambda}$ is a true norm and

$$\|\mathbf{P}_k - \mathbf{Q}_k\|^2_{\xi,C_\lambda} \to 0 \implies \mathbf{P}_k, \mathbf{Q}_k \to \tilde{\mathbf{P}}.$$

When $\lambda = 0$ we have no guarantees on what happens to the e component of either $\mathbf{P}_k$ or $\mathbf{Q}_k$, and we can only say that $\mathbf{P}_k$ (resp., $\mathbf{Q}_k$) converges to a set S whose elements differ by a constant component.

Using a variation on a standard argument (Tsitsiklis & Van Roy, 1997), we now write (in ξ-weighted norm)

$$\begin{aligned}
l^2_{\xi,\lambda}(\tilde{\mathbf{P}}, \mathbf{P}^\pi) &= l^2_{\xi,\lambda}(\hat{\mathbf{\Pi}}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\tilde{\mathbf{P}}, \mathbf{P}^\pi) \\
&\qquad\qquad \text{(By definition of } \tilde{\mathbf{P}}\text{)} \\
&= l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}}, \mathbf{P}^\pi) \\
&= l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}}, \mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi) \\
&\quad + l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi, \mathbf{P}^\pi) \\
&\qquad\qquad \text{(Using Eq. (11))} \\
&= l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}}, \mathbf{\Pi}_{\xi,\lambda,\Phi}\mathcal{T}^\pi\mathbf{P}^\pi) \\
&\quad + l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi, \mathbf{P}^\pi) \\
&\qquad\qquad (\mathbf{P}^\pi \text{ is the fixed point of } \mathcal{T}^\pi) \\
&\le l^2_{\xi,\lambda}(\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}}, \mathcal{T}^\pi\mathbf{P}^\pi) + l^2_{\xi,\lambda}(\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi, \mathbf{P}^\pi).
\end{aligned}$$

We now focus on the first term. Unlike Tsitsiklis & Van Roy (1997)'s argument, we are faced here with two different operators: $\tilde{\mathcal{T}}^\pi$ and $\mathcal{T}^\pi$. We write

$$\begin{aligned}
l^2_{\xi,\lambda}(\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}}, \mathcal{T}^\pi\mathbf{P}^\pi) &= \left\|\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}} - \mathcal{T}^\pi\mathbf{P}^\pi\right\|^2_{\xi,C_\lambda} \\
&= \left\|\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}} - \mathcal{T}^\pi\mathbf{P}^\pi\right\|^2_{\xi,AA^\top} \\
&\quad + \lambda\left\|\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}} - \mathcal{T}^\pi\mathbf{P}^\pi\right\|^2_{\xi,ee^\top}.
\end{aligned}$$

Because $\tilde{\mathcal{T}}^\pi$ "resets" the distribution's mass to 1, the second term is zero. Similarly,

$$\begin{aligned}
\left\|\tilde{\mathcal{T}}^\pi\tilde{\mathbf{P}} - \mathcal{T}^\pi\mathbf{P}^\pi\right\|^2_{\xi,AA^\top} &= \left\|\mathcal{T}^\pi\tilde{\mathbf{P}} - \mathcal{T}^\pi\mathbf{P}^\pi\right\|^2_{\xi,AA^\top} \\
&\le \gamma\left\|\tilde{\mathbf{P}} - \mathbf{P}^\pi\right\|^2_{\xi,AA^\top} \\
&= \gamma\left\|\tilde{\mathbf{P}} - \mathbf{P}^\pi\right\|^2_{\xi,C_\lambda} \\
&\quad - \gamma\lambda\left\|\tilde{\mathbf{P}} - \mathbf{P}^\pi\right\|^2_{\xi,ee^\top}.
\end{aligned}$$

Expanding the first inequality repeatedly, we put everything together and find that

$$\begin{aligned}
l^2_{\xi,\lambda}(\tilde{\mathbf{P}}, \mathbf{P}^\pi) &\le \frac{1}{1-\gamma}\|\mathbf{\Pi}_{\xi,\lambda,\Phi}\mathbf{P}^\pi - \mathbf{P}^\pi\|^2_{\xi,C_\lambda} \\
&\quad - \frac{\gamma\lambda}{1-\gamma}\left\|\tilde{\mathbf{P}} - \mathbf{P}^\pi\right\|^2_{\xi,ee^\top}. \qquad \square
\end{aligned}$$

Corollary 1. *Under the same conditions as those used by Tsitsiklis & Van Roy (1997), the stochastic update process where one samples $x \sim \xi$ and updates the parameter Θ according to*

$$\Theta_{k+1} \leftarrow \Theta_k + \alpha_k \nabla_\Theta l^2_\lambda(\hat{\mathcal{T}}^\pi\mathbf{P}_k(x), \mathbf{P}_k(x)),$$

where $\hat{\mathcal{T}}^\pi$ is the random operator derived from a sample transition (x, r, x'), also converges.

To prove Theorem 2, we will need the following result:

Lemma 7 (Ratio of operators)**.** *Let M be a self-adjoint linear operator and N be a self-adjoint, invertible linear operator. Then*

$$\sup_f \frac{< f, Mf >}{< f, Nf >} = \rho\left(N^{-1/2}MN^{-1/2}\right),$$

where $\rho(\cdot)$ denotes the spectral radius of its argument.

Proof. Denoting $g = N^{1/2}f$, we have

$$\begin{aligned}
f &= N^{-1/2}g \\
\frac{< f, Mf >}{< f, Nf >} &= \frac{< N^{-1/2}g, MN^{-1/2}g >}{< g, g >} \\
&= \frac{< g, N^{-1/2}MN^{-1/2}g >}{< g, g >}.
\end{aligned}$$

Taking the supremum over g gives the desired result. □

Theorem 2 (Error bound for the expected value). *Let $\|\cdot\|_\xi$ be the ξ-weighted norm over value functions. The squared expectation error of the fixed point $\tilde{\mathbf{P}}$ with respect to the true value function V^π is bounded as*

$$\|\mathbb{E}_{\tilde{\mathbf{P}}}\,\mathbf{z} - V^\pi\|_\xi^2 \leq \|C_\lambda^{-1/2}\mathbf{z}\|^2 \boldsymbol{l}_{\xi,\lambda}^2(\tilde{\mathbf{P}}, \mathbf{P}^\pi).$$

Proof.

$$\begin{aligned}
&\|\mathbb{E}_{\tilde{\mathbf{P}}}\,\mathbf{z} - V^\pi\|_\xi^2 = \\
&\quad= \sum_{x\in\mathcal{X}} \xi(x)\langle \tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x), \mathbf{z}\mathbf{z}^*\big(\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)\big)\rangle \\
&\quad= \sum_{x\in\mathcal{X}} \xi(x)\langle \tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x), C_\lambda\big(\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)\big)\rangle \\
&\qquad\quad \times \frac{\langle \tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x), \mathbf{z}\mathbf{z}^*\big(\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)\big)\rangle}{\langle \tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x), C_\lambda\big(\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)\big)\rangle} \\
&\quad\leq \sum_{x\in\mathcal{X}} \xi(x)\langle \tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x), C_\lambda\big(\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)\big)\rangle \\
&\qquad\quad \times \max_f \frac{\langle f, \mathbf{z}\mathbf{z}^* f\rangle}{\langle f, C_\lambda f\rangle} \\
&\quad\overset{(a)}{=} \sum_{x\in\mathcal{X}} \xi(x)\langle \tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x), C_\lambda\big(\tilde{\mathbf{P}}(x) - \mathbf{P}^\pi(x)\big)\rangle \\
&\qquad\quad \times \|C_\lambda^{-1/2}\mathbf{z}\|^2 \\
&\quad= \|C_\lambda^{-1/2}\mathbf{z}\|^2 l_{\xi,\lambda}^2(\boldsymbol{\Pi}_{\xi,\lambda,\Phi}^\pi, \mathbf{P}^\pi)
\end{aligned}$$

where the step a) uses Lemma 7 and the fact that $A^{-1/2}\mathbf{z}\mathbf{z}^* A^{-1/2}$ is a rank one operator. □

A Experimental Details

Our S51 implementation is based on the C51 code from the Dopamine framework Castro et al. (2018), with only minor modifications to account for the new loss. Specifically, we

1. Remove the softmax transfer function mapping logits to probabilities; our network's outputs $o(x, a)$ are directly used as "probabilities";
2. Select actions according to the maximum predicted "expectation", which is $\mathbf{z}^\top o(x, a)$, where $\mathbf{z}$ is a 51-dimensional vector whose entries are uniformly spaced within $[-10, 10]$;
3. Replace the cross-entropy loss by the modified squared loss defined in Equation 9.

For C51, we used the hyperparameters provided by Bellemare et al. (2017a). We optimized the hyperparameters for S51 over the same range as used in that paper, and found that a smaller step size ($\alpha = 2.5 \times 10^{-5}$, vs. 2.5×10^{-4} for C51) and optimizer epsilon ($\epsilon_{\text{OPT}} = 3.125 \times 10^{-5}$, vs 3.125×10^{-4}) performed best. The parameter $\lambda = 10$ was selected from a hyperparameter sweep ($\lambda \in \{0, 0.25, 1, 10, 20, 100\}$); we found the method to perform reasonably the same for a broad range of λ values, but note that $\lambda = 0$ yielded worse performance. In both cases, the training epsilon was set to $\epsilon = 0.05$, and lives lost were counted as the end of an episode.

games	**video url**
Asterix	`https://youtu.be/hk4sYkx-VuQ`
Breakout	`https://youtu.be/POWvu9-2m6E`
Pong	`https://youtu.be/f63K_peZ6uE`
Seaquest	`https://youtu.be/lbySDvtAmPo`
Space Invaders	`https://youtu.be/dMvN9gmAy7E`

Figure 3: Links to videos of the S51 value distributions after training.