

Non-parametric Approximate Dynamic Programming via the Kernel Method

A . Online Supplement

A .1. Introduction

This paper supplements our submission in NIPS 2012. The supplement aids the paper in the following ways:

- It provides technical proofs that are not necessary for the conceptual understanding of the paper.
- It outlines the numerical scheme we use to solve the optimization problem of interest.

Section A .2 provies the proof of Proposition 1 of the NIPS 2012 paper. Section A .3 proves the main theoretical guarantee. Finally, Section A .4 outlines and proves the correctness of the numerical procedure we use to solve the program of interest.

A .2. Duality of the Sampled RSALP

We aim to approximate the optimal cost-to-go function via solving the sampled version of the Regularized Smoothed Approximate Linear Program (RSALP):

$$\begin{aligned}
 (1) \quad & \text{maximize} \quad \frac{1}{N} \sum_{x \in \hat{\mathcal{S}}} w_x \langle \mathbf{x}, \mathbf{z} \rangle + b - \frac{\kappa}{N} \sum_{x \in \hat{\mathcal{S}}} s_x - \frac{\Gamma}{2} \langle \mathbf{z}, \mathbf{z} \rangle \\
 & \text{subject to} \quad \langle \mathbf{x}, \mathbf{z} \rangle + b \leq g_{a,x} + \alpha \mathbf{E}_{x,a}[\langle \mathbf{X}', \mathbf{z} \rangle + b] + s_x, \quad \forall x \in \hat{\mathcal{S}}, a \in \mathcal{A}, \\
 & \quad \quad \quad s_x \geq 0, \quad \quad \quad \forall x \in \hat{\mathcal{S}}, \\
 & \quad \quad \quad \mathbf{z} \in \mathcal{H}, b \in \mathbb{R}, s \in \mathbb{R}^{\hat{\mathcal{S}}}.
 \end{aligned}$$

We do so by numerically solving its dual¹,

$$\begin{aligned}
 (2) \quad & \text{minimize} \quad \frac{1}{2} \lambda^\top Q \lambda + R^\top \lambda + S \\
 & \text{subject to} \quad \sum_{a \in \mathcal{A}} \lambda_{x,a} \leq \frac{\kappa}{N} \quad \forall x \in \hat{\mathcal{S}}, \\
 & \quad \quad \quad \sum_{x \in \hat{\mathcal{S}}, a \in \mathcal{A}} \lambda_{x,a} = \frac{1}{1 - \alpha}, \lambda \geq 0, \\
 & \quad \quad \quad \lambda \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}.
 \end{aligned}$$

In the main paper we have the dual program without the offset S . Here we define S as:

$$S \triangleq - \sum_{x \in \hat{\mathcal{S}}} \sum_{y \in \hat{\mathcal{S}}} w_x w_y \langle \mathbf{x}, \mathbf{y} \rangle.$$

As we'll soon see, adding this constant makes the primal and dual optima equal.

¹To be made precise in Proposition 1

Proposition 1. *Programs (1) and (2) have equal (finite) optimal values. The optimal solution to (2) is attained at some λ^* . The optimal solution to (1) is attained at some (z^*, s^*, b^*) with*

$$(3) \quad \mathbf{z}^* = \frac{1}{\Gamma} \left[\frac{1}{N} \sum_{x \in \hat{\mathcal{S}}} w_x \mathbf{x} - \sum_{x \in \hat{\mathcal{S}}, a \in \mathcal{A}} \lambda_{x,a}^* \left(\mathbf{x} - \alpha \mathbf{E}_{x,a}[\mathbf{X}'] \right) \right].$$

Proof. We begin with a few observations about the primal program, (1):

1. The weight vector, \mathbf{z} can be restricted without loss to some finite ball in \mathcal{H} . The optimal value of the primal is consequently finite.
2. The primal has a feasible interior point: consider setting $\mathbf{z} = \mathbf{0}$, $b = 0$, and $s_x = \max(-\min_a g_{x,a}, \epsilon)$ for some $\epsilon > 0$.
3. The optimal value of the primal is achieved. To see this, we note that it suffices to restrict \mathbf{z} to the finite dimensional space spanned by the vectors in $\{\mathbf{x} : x \in \hat{\mathcal{S}}\}$, so that the feasible region of the primal can now be restricted, without loss of optimality, to a compact subset of $\mathcal{H} \times \mathbb{R}^{|\hat{\mathcal{S}}|} \times \mathbb{R}$. Since the objective function of the primal is continuous, we know that its optimal value must be achieved by the Weierstrass theorem.

We next derive the dual to (1). As in Luenberger [1997], Chapter 8, we define the Lagrangian:

$$\begin{aligned} \mathcal{L}(\mathbf{z}, b, s, \lambda) \triangleq & \left\langle -\frac{1}{N} \sum_{x \in \hat{\mathcal{S}}} w_x \mathbf{x} + \sum_{(x,a) \in \hat{\mathcal{S}} \times \mathcal{A}} \lambda_{x,a} (\mathbf{x} - \alpha \mathbf{E}_{x,a}[\mathbf{X}']), \mathbf{z} \right\rangle + \frac{C}{2} \|\mathbf{z}\|_{\mathcal{H}}^2 \\ & + \sum_{x \in \hat{\mathcal{S}}} s_x \left(\frac{\kappa}{N} - \sum_{a \in \mathcal{A}} \lambda_{x,a} \right) - b \left(1 - (1 - \alpha) \sum_{(x,a) \in \hat{\mathcal{S}} \times \mathcal{A}} \lambda_{x,a} \right) - \sum_{(x,a) \in \hat{\mathcal{P}}} g_{x,a} \lambda_{x,a}. \end{aligned}$$

and define the dual function $G(\lambda) \triangleq \inf_{(\mathbf{z}, b, s) \in \mathcal{D}} \mathcal{L}(\mathbf{z}, b, s, \lambda)$ where we denote by \mathcal{D} the feasible region of the primal problem. Now, observe that for any given λ , $\mathcal{L}(\mathbf{z}, b, s, \lambda)$ is (uniquely) minimized at

$$(4) \quad \mathbf{z}^*(\lambda) = \frac{1}{\Gamma} \left[\frac{1}{N} \sum_{x \in \hat{\mathcal{S}}} w_x \mathbf{x} - \sum_{x \in \hat{\mathcal{S}}, a \in \mathcal{A}} \lambda_{x,a} \left(\mathbf{x} - \alpha \mathbf{E}_{x,a}[\mathbf{X}'] \right) \right],$$

for any finite b, s . This follows from the observation that for any $\bar{\mathbf{z}} \in \mathcal{H}$, $\langle \mathbf{z}, \mathbf{z} \rangle - \langle \bar{\mathbf{z}}, \mathbf{z} \rangle$ is minimized at $-\frac{1}{2}\bar{\mathbf{z}}$ by the Cauchy-Schwarz inequality. It follows immediately that on the set defining the feasible region of the program (2), we must have that

$$G(\lambda) = \frac{1}{2} \lambda^\top Q \lambda + R^\top \lambda + S$$

and moreover that $G(\lambda) = +\infty$ outside that set. This suffice to establish that the dual problem $\inf_{\lambda \geq 0} G(\lambda)$ is precisely program (2).

The first conclusion of Theorem 1, pp. 224-225 of Luenberger [1997] and the first and second observations we made at the outset of our proof then suffice to establish that programs (1) and (2) have equal optimal values (i.e. strong duality holds) and that the optimal value of the dual program is achieved at some λ^* . Our third observation, (4), and the second conclusion of Theorem 1, pp. 224-225 of Luenberger [1997] then suffice to establish our second claim. ■

A .3. Proof of the Performance Guarantee

A .3.1. Guarantee for the Exact Program

In this section we will prove a stronger guarantee than Theorem 1 of the NIPS2012 submission. Before proceeding with our analysis, we introduce some additional notation. Let $\Psi = \{\psi \in \mathbb{R}^{\mathcal{S}} : \psi \geq \mathbf{1}\}$, be the set of all functions on the state space bounded from below by 1. For any $\psi \in \Psi$, let us define the weighted ∞ -norm $\|\cdot\|_{\infty, 1/\psi}$ by

$$\|J\|_{\infty, 1/\psi} \triangleq \max_{x \in \mathcal{S}} \frac{|J(x)|}{\psi(x)}.$$

Our use of such weighted norms will allow us to emphasize approximation error differently across the state space. Further, we define

$$\beta(\psi) \triangleq \max_{x, a} \frac{\sum_{x'} p(x, x', a) \psi(x')}{\psi(x)}.$$

For a given ψ , $\beta(\psi)$ gives us the worst-case expected gain of the Lyapunov function ψ for any state action pair (x, a) .

Let us redefine the ‘idealized’ distribution. Let ν be an arbitrary distribution over \mathcal{S} and denote by μ^* and P_{μ^*} an optimal policy and the transition probability matrix corresponding to this policy respectively. We then define a distribution over \mathcal{S} , $\pi_{\mu^*, \nu}$ according to

$$(5) \quad \pi_{\mu^*, \nu}^\top \triangleq (1 - \alpha) \nu^\top (I - \alpha P_{\mu^*})^{-1} = (1 - \alpha) \sum_{t=0}^{\infty} \alpha^t \nu^\top P_{\mu^*}^t.$$

This idealized distribution will play the role of the sampling distribution π in the sequel.

It is notationally convenient for us to introduce the Bellman operator defined by

$$(TJ)(x) = \min_{a \in \mathcal{A}} \left[g(x, a) + \alpha \mathbb{E}_{x, a} [J(X')] \right],$$

for all $x \in \mathcal{S}$. Let $\hat{\mathcal{S}}$ be a set of N states drawn independently at random from \mathcal{S} ; here we pick a specific sampling distribution $\pi = \pi_{\mu^*, \nu}$. The ‘idealized’ sampled program we consider is:

$$(6) \quad \begin{aligned} & \text{maximize} \quad \nu^\top \tilde{J}_{\mathbf{z}, b} - \frac{2}{1-\alpha} \frac{1}{N} \sum_{x \in \hat{\mathcal{S}}} s_x \\ & \text{subject to} \quad \tilde{J}_{\mathbf{z}, b} \leq T \tilde{J}_{\mathbf{z}, b} + s, \\ & \quad \|\mathbf{z}\|_{\mathcal{H}} \leq C, \quad |b| \leq B, \quad s \geq 0, \\ & \quad \mathbf{z} \in \mathcal{H}, \quad b \in \mathbb{R}, \quad s \in \mathbb{R}^{\hat{\mathcal{S}}}. \end{aligned}$$

Let $(\hat{\mathbf{z}}, \hat{b})$ be an optimal solution to the idealized sampled program, (6) and let $K \triangleq \max_{x \in \mathcal{S}} \|\mathbf{x}\|_{\mathcal{H}}$. Further define:

$$\Xi(C, B, K, \delta) \triangleq \left(4CK(1 + \alpha) + 4B(1 - \alpha) + 2\|g\|_{\infty} \right) \left(1 + \sqrt{\frac{1}{2} \ln(1/\delta)} \right).$$

Notice that $\Xi(C, B, K, \delta)^2$ scales as the square of C, K and B and further is $O(\ln 1/\delta)$. The following result will constitute our main approximation guarantee:

Theorem 1. For any $\epsilon, \delta > 0$, let $N \geq \frac{\Xi(C, B, K, \delta)^2}{\epsilon^2}$. If $(\hat{\mathbf{z}}, \hat{b})$ is an optimal solution to (6), then with probability at least $1 - \delta - \delta^4$

$$(7) \quad \|J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}\|_{1, \nu} \leq \inf_{\|\mathbf{z}\|_{\mathcal{H}} \leq C, |b| \leq B, \psi \in \Psi} \|J^* - \tilde{J}_{\mathbf{z}, b}\|_{\infty, 1/\psi} \left(\nu^\top \psi + \frac{2(\pi_{\mu^*, \nu}^\top \psi)(\alpha\beta(\psi) + 1)}{1 - \alpha} \right) + \frac{4\epsilon}{1 - \alpha}.$$

We prepare the ground for the proof by developing appropriate uniform concentration guarantees for appropriate function classes.

A .3.2. Uniform Concentration Bounds

We begin with defining the empirical Rademacher complexity of a class of functions \mathcal{F} from \mathcal{S} to \mathbb{R} as

$$\hat{R}_n(\mathcal{F}) = \mathbb{E} \left[\sup_{f \in \mathcal{F}} \frac{2}{n} \sum_{i=1}^n \sigma_i f(X_i) \middle| X_1, X_2, \dots, X_n \right]$$

where σ_i are i.i.d. random variable that take value 1 with probability 1/2 and -1 with probability 1/2. The X_i are i.i.d. \mathcal{S} valued random variables drawn with the distribution π . We denote by $R_n(\mathcal{F}) = \mathbb{E} \hat{R}_n(\mathcal{F})$ the Rademacher complexity of \mathcal{F} .

We begin with the following abstract sample complexity result: Let \mathcal{F} be a class of functions mapping \mathcal{S} to \mathbb{R} that are uniformly bounded by some constant \bar{B} . Moreover denote for any function $f \in \mathcal{F}$, the empirical expectation $\hat{\mathbb{E}}_n f(X) \triangleq \frac{1}{n} \sum_{i=1}^n f(X_i)$ where the X_i are i.i.d. We then have the following sample complexity result:

Lemma 1.

$$\mathbb{P} \left(\sup_{f \in \mathcal{F}} \mathbb{E} f(X) - \hat{\mathbb{E}}_n f(X) \geq R_n(\mathcal{F}) + \sqrt{\frac{2\bar{B}^2 \ln(1/\delta)}{n}} \right) \leq \delta.$$

This result is standard; for completeness, the proof may be found in Appendix ???. Next, we establish the Rademacher complexity of a specific class of functions. Fixing a policy μ , consider then the set of functions mapping \mathcal{S} to \mathbb{R} defined according to:

$$\mathcal{F}_{\mathcal{S}, \mu} \triangleq \left\{ x \mapsto \langle \mathbf{x}, \mathbf{z} \rangle - \alpha \mathbb{E}_{x, \mu(x)}[\langle \mathbf{X}', \mathbf{z} \rangle] : \|\mathbf{z}\|_{\mathcal{H}} \leq C \right\}$$

We have:

Lemma 2. For any policy μ ,

$$R_n(\mathcal{F}_{\mathcal{S}, \mu}) \leq \frac{2CK(1 + \alpha)}{\sqrt{n}}.$$

Proof. Observe that, due to triangle inequality

$$\|\mathbf{x} - \alpha \mathbb{E}_{x, \mu(x)}[\mathbf{X}']\|_{\mathcal{H}} \leq \|\mathbf{x}\|_{\mathcal{H}} + \alpha \mathbb{E}_{x, \mu(x)}[\|\mathbf{X}'\|_{\mathcal{H}}] \leq K(1 + \alpha),$$

for all $x \in \mathcal{S}$. Now, let X_i be i.i.d. samples in \mathcal{S} and \mathbf{X}_i be the corresponding elements in \mathcal{H} ,

$$\begin{aligned}
\hat{R}_n(\mathcal{F}_{\mathcal{S},\mu}) &= \frac{2}{n} \mathbb{E} \left[\sup_{z: \|z\|_{\mathcal{H}} \leq C} \left\langle \sum_i \sigma_i \left(\mathbf{X}_i - \alpha \mathbb{E}_{X_i, \mu(X_i)}[\mathbf{X}'] \right), z \right\rangle \middle| X_1, \dots, X_n \right] \\
&\leq \frac{2}{n} \mathbb{E} \left[\sup_{z: \|z\|_{\mathcal{H}} \leq C} \left\| \sum_i \sigma_i \left(\mathbf{X}_i - \alpha \mathbb{E}_{X_i, \mu(X_i)}[\mathbf{X}'] \right) \right\|_{\mathcal{H}} \|z\|_{\mathcal{H}} \middle| X_1, \dots, X_n \right] \\
&= \frac{2C}{n} \mathbb{E} \left[\left\| \sum_i \sigma_i \left(\mathbf{X}_i - \alpha \mathbb{E}_{X_i, \mu(X_i)}[\mathbf{X}'] \right) \right\|_{\mathcal{H}} \middle| X_1, \dots, X_n \right] \\
&\leq \frac{2C}{n} \sqrt{\sum_i \|\mathbf{X}_i - \alpha \mathbb{E}_{X_i, \mu(X_i)}[\mathbf{X}']\|_{\mathcal{H}}^2} \\
&\leq \frac{2CK(1+\alpha)}{\sqrt{n}}.
\end{aligned}$$

■

Now, consider the class of functions mapping \mathcal{S} to \mathbb{R} , defined according to:

$$\overline{\mathcal{F}}_{\mathcal{S},\mu} \triangleq \left\{ x \mapsto \left(\tilde{J}_{\mathbf{z},b}(x) - (T_{\mu} \tilde{J}_{\mathbf{z},b})(x) \right)^+ : \|\mathbf{z}\|_{\mathcal{H}} \leq C, b \leq B \right\}$$

Now, $\overline{\mathcal{F}}_{\mathcal{S},\mu} = \phi \cdot (\mathcal{F}_{\mathcal{S},\mu} + (1-\alpha)\mathcal{F}_B - g_{\mu})$, where $\phi = (\cdot)^+$ and $\mathcal{F}_B = \{x \mapsto b : |b| \leq B\}$. It is easy to show that $R_n(\mathcal{F}_B) \leq 2B/\sqrt{n}$, so that with the previous Lemma, Theorem 12, parts 4 and 5 of Bartlett and Mendelson [2002] allow us to conclude

Corollary 1.

$$R_n(\overline{\mathcal{F}}_{\mathcal{S},\mu}) \leq \frac{4CK(1+\alpha) + 4B(1-\alpha) + 2\|g_{\mu}\|_{\infty}}{\sqrt{n}} \triangleq \frac{\overline{C}}{\sqrt{n}}$$

Now, define

$$\overline{\mathcal{F}}_{\mathcal{S}} \triangleq \left\{ x \mapsto \left(\tilde{J}_{\mathbf{z},b}(x) - (T \tilde{J}_{\mathbf{z},b})(x) \right)^+ : \|\mathbf{z}\|_{\mathcal{H}} \leq C, b \leq B \right\}$$

We have:

Lemma 3. *For every $f \in \overline{\mathcal{F}}_{\mathcal{S}}$ we have that $\|f\|_{\infty} \leq \overline{C}/2$. Moreover,*

$$\mathbb{P} \left(\hat{\mathbb{E}}_N f(X) - \mathbb{E} f(X) \geq \epsilon \right) \leq \delta^4$$

provided $N \geq \frac{\Xi(C,B,K,\delta)^2}{\epsilon^2}$.

The first claim above follows from routine algebra and the Cauchy-Schwarz inequality; the second is Hoeffding's inequality. Corollary 1, Lemma 1 and the first part of Lemma 3 yields the following sample complexity result:

Theorem 2. *Provided*

$$N \geq \frac{\Xi(C,B,K,\delta)^2}{\epsilon^2}$$

we have

$$\mathbb{P} \left(\sup_{f \in \overline{\mathcal{F}}_{\mathcal{S},\mu}} \mathbb{E} f(X) - \hat{\mathbb{E}}_N f(X) \geq \epsilon \right) \leq \delta.$$

Theorem 2 will constitute a crucial sample complexity bound for our main result; we now proceed with the proof of Theorem 1.

A.3.3. Proof of Theorem 1

Let $(\hat{\mathbf{z}}, \hat{b}, \hat{s})$ be the optimal solution to the sampled program (6). Define

$$\hat{s}_{\mu^*} = (\tilde{J}_{\hat{\mathbf{z}}, \hat{b}} - T_{\mu^*} \tilde{J}_{\hat{\mathbf{z}}, \hat{b}})^+.$$

Observe that we may assume without loss of generality that

$$\hat{s} = (\tilde{J}_{\hat{\mathbf{z}}, \hat{b}} - T \tilde{J}_{\hat{\mathbf{z}}, \hat{b}})^+,$$

so that $\hat{s} \geq \hat{s}_{\mu^*}$. Now, by definition, $\tilde{J}_{\hat{\mathbf{z}}, \hat{b}} \leq T_{\mu^*} \tilde{J}_{\hat{\mathbf{z}}, \hat{b}} + \hat{s}_{\mu^*}$, so that by Lemma 2 of Desai et al. [2011], we have that

$$\tilde{J}_{\hat{\mathbf{z}}, \hat{b}} \leq J^* + \Delta^* \hat{s}_{\mu^*},$$

where $\Delta^* = (I - \alpha P_{\mu^*})^{-1}$. Let $\hat{\pi}_{\mu^*, \nu}$ be the empirical distribution obtained by sampling N states according to $\pi_{\mu^*, \nu}$. Now let \mathbf{z}, b satisfying $\|\mathbf{z}\|_{\mathcal{H}} \leq C, |b| \leq B$ be given, and define $s_{\mathbf{z}, b} \triangleq (T \tilde{J}_{\mathbf{z}, b} - \tilde{J}_{\mathbf{z}, b})^+$. Then, $(\mathbf{z}, b, s_{\mathbf{z}, b})$ constitute a feasible solution to (6). Finally, let $\psi \in \Psi$ be arbitrary. We then have with probability at least $1 - \delta - \delta^4$,

$$\begin{aligned}
 \|J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}\|_{1, \nu} &= \|J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}} + \Delta^* \hat{s}_{\mu^*} - \Delta^* \hat{s}_{\mu^*}\|_{1, \nu} \\
 &\leq \|J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}} + \Delta^* \hat{s}_{\mu^*}\|_{1, \nu} + \|\Delta^* \hat{s}_{\mu^*}\|_{1, \nu} \\
 &= \nu^\top (J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}) + 2\nu^\top \Delta^* \hat{s}_{\mu^*} \\
 &= \nu^\top (J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}) + \frac{2}{1 - \alpha} \pi_{\mu^*, \nu}^\top \hat{s}_{\mu^*} \\
 &\leq \nu^\top (J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}) + \frac{2}{1 - \alpha} \hat{\pi}_{\mu^*, \nu}^\top \hat{s}_{\mu^*} + \frac{2\epsilon}{1 - \alpha} \\
 &\leq \nu^\top (J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}) + \frac{2}{1 - \alpha} \hat{\pi}_{\mu^*, \nu}^\top \hat{s} + \frac{2\epsilon}{1 - \alpha} \\
 &\leq \nu^\top (J^* - \tilde{J}_{\mathbf{z}, b}) + \frac{2}{1 - \alpha} \hat{\pi}_{\mu^*, \nu}^\top s_{\mathbf{z}, b} + \frac{2\epsilon}{1 - \alpha} \\
 &\leq \nu^\top (J^* - \tilde{J}_{\mathbf{z}, b}) + \frac{2}{1 - \alpha} \pi_{\mu^*, \nu}^\top s_{\mathbf{z}, b} + \frac{4\epsilon}{1 - \alpha} \\
 &\leq (\nu^\top \psi) \|J^* - \tilde{J}_{\mathbf{z}, b}\|_{\infty, 1/\psi} + \frac{2}{1 - \alpha} (\pi_{\mu^*, \nu}^\top \psi) \|T \tilde{J}_{\mathbf{z}, b} - \tilde{J}_{\mathbf{z}, b}\|_{\infty, 1/\psi} + \frac{4\epsilon}{1 - \alpha}.
 \end{aligned}
 \tag{8}$$

The second equality follows by our observation that $\tilde{J}_{\hat{\mathbf{z}}, \hat{b}} \leq J^* + \Delta^* \hat{s}_{\mu^*}$ and since $\Delta^* \hat{s}_{\mu^*} \geq 0$. The second inequality above holds with probability at least $1 - \delta$ by virtue of Theorem 2 and the fact that $\hat{s}_{\mu^*} \in \overline{\mathcal{F}}_{\mathcal{S}, \mu^*}$. The subsequent inequality follows from our observation that $\hat{s} \geq \hat{s}_{\mu^*}$. The fourth inequality follows from the assumed optimality of $(\hat{\mathbf{z}}, \hat{b}, \hat{s})$ for the sampled program (6) and the feasibility of $(\mathbf{z}, b, s_{\mathbf{z}, b})$ for the same. The fifth inequality holds with probability $1 - \delta^4$ and follows from the Hoeffding bound in Lemma 3 since $s_{\mathbf{z}, b} \in \overline{\mathcal{F}}_{\mathcal{S}}$. The final inequality follows from the observation that for any $s \in \mathbb{R}^{\mathcal{S}}, \psi \in \Psi$ a and probability vector ν , $\nu^\top s \leq (\nu^\top \psi) \|s\|_{\infty, 1/\psi}$.

Now the proof of Theorem 2 in [Desai et al., 2011] establishes that for any $\psi \in \Psi$ and $J \in \mathbb{R}^{\mathcal{S}}$, we have,

$$\|TJ - J\|_{\infty, 1/\psi} \leq (1 + \alpha\beta(\psi)) \|J^* - J\|_{\infty, 1/\psi}.$$

Applied to (8), this yields

$$\|J^* - \tilde{J}_{\hat{\mathbf{z}}, \hat{b}}\|_{1, \nu} \leq \|J^* - \tilde{J}_{\mathbf{z}, b}\|_{\infty, 1/\psi} \left(\nu^\top \psi + \frac{2(\pi_{\mu^*, \nu}^\top \psi)(\alpha\beta(\psi) + 1)}{1 - \alpha} \right) + \frac{4\epsilon}{1 - \alpha}.$$

Since our choice of \mathbf{z}, b was arbitrary (beyond satisfying $\|\mathbf{z}\|_{\mathcal{H}} \leq C, |b| \leq B$), the result follows.

A .4. Numerical Scheme

This section outlines an efficient numerical scheme we use to solve the regularized SALP. In particular, we would like to solve the sampled dual problem (2), in order to find an approximation to the optimal cost-to-go function. This approach requires solving a QP with the number of variables equal to NA , where $N \triangleq |\hat{\mathcal{S}}|$ is the number of sampled states and $A \triangleq |\mathcal{A}|$ is the number of possible actions. Furthermore, constructing the coefficient matrices Q and R for (2) requires numerical effort of the order $O(N^2 A^2 H^2)$, where H is the maximum number of states that can be reached from an arbitrary state-action pair, i.e.,

$$H \triangleq \max_{(x,a) \in \hat{\mathcal{S}} \times \mathcal{A}} |\{x' \in \mathcal{S} \mid p(x, x', a) > 0\}|.$$

These costly steps may prevent scaling up solution of the QP to a large number of samples. Also, an off-the-shelf QP solver will also require storing the matrix Q in memory, hence the memory requirement scales as $O(N^2 A^2)$.

In this section, the problem structure of (2) is exploited to design an iterative numerical scheme that requires per iteration numerical effort and memory requirement that scales at a substantially lower rate. The method is similar to the efficient implementation of SVM solvers discussed by Osuna et al. [1997] or Joachims [1999], for example. The main idea is to employ an active set method, where all but a small subset of variables are frozen and the corresponding subproblem is solved. In each iteration t , we compute an estimate of $\lambda^{(t)} \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ of an optimal solution to (2). The decision variables are updated in an iterative fashion. Given the prior estimate $\lambda^{(t-1)}$, the updated solution estimate $\lambda^{(t)}$ is chosen to be a feasible vector for (2), and is computed as follows:

1. A subset $\mathcal{B} \subset \hat{\mathcal{S}} \times \mathcal{A}$ of the decision variables in (2) is chosen. The size of this subset will be much smaller than the total number of decision variables, in fact, we shall see that subsets of cardinality $|\mathcal{B}| = 2$ are sufficient. If a subset \mathcal{B} of low cardinality that contains a descent direction cannot be found, we will establish that the prior iterate $\lambda^{(t-1)}$ is, in fact, a globally optimal solution to the original problem (2).
2. Given the subset \mathcal{B} , we solve (2) for $\lambda^{(t)}$, where all indices except those in \mathcal{B} are frozen at their prior values. In other words, $\lambda_{x,a}^{(t)} \triangleq \lambda_{x,a}^{(t-1)}$ for all $(x,a) \notin \mathcal{B}$. This entails the solution of a QP with only $|\mathcal{B}|$ decision variables.

In the following sections, we will discuss the algorithmic details of each step in our iterative method, as well as establishing correctness of the overall procedure: each iteration maintains feasibility while decreasing the objective value of (2), and the procedure will only terminate when a globally optimal iterate has been reached. We will establish that, in each of the above steps, the total amount of work is at most $O(NA \log NA)$. Moreover, we shall see that our iterative procedure does not require storage of the entire matrix Q , and has memory requirements of order $O(NA)$.

A .4.1. Subset Selection

The first step in the active set method is to choose the subset $\mathcal{B} \subset \hat{\mathcal{S}} \times \mathcal{A}$ of decision variables to optimize over. Given the convex objective in (2), if the prior iteration of the algorithm is at a suboptimal point $\lambda \triangleq \lambda^{(t-1)}$, then there exists a direction $d \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ such that $\lambda + \epsilon d$ is feasible with a lower objective value for $\epsilon > 0$ sufficiently small. To select a subset to optimize over, we look for such a descent direction d of low cardinality $\|d_0\|_0 \leq q$, i.e., a vector d that is zero on all but at most q components. If such direction can be found, then we can use the set of non-zero indices of d as our set \mathcal{B} .

This problem of finding a descent direction d can be posed as

$$\begin{aligned}
(9) \quad & \text{minimize} && h(\lambda)^\top d \\
& \text{subject to} && \sum_{a \in \mathcal{A}} d_{x,a} \leq 0, \quad \forall x \in \hat{\mathcal{S}} \text{ with } \sum_{x \in \mathcal{A}} \lambda_{x,a} = \frac{\kappa}{N}, \\
& && d_{x,a} \geq 0, \quad \forall (x,a) \in \hat{\mathcal{S}} \times \mathcal{A} \text{ with } \lambda_{x,a} = 0, \\
& && \sum_{\substack{x \in \hat{\mathcal{S}} \\ a \in \mathcal{A}}} d_{x,a} = 0, \\
& && \|d\|_0 \leq q, \\
& && \|d\|_\infty \leq 1, \\
& && d \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}.
\end{aligned}$$

Here, $h(\lambda) \triangleq Q\lambda + R$ is the gradient of the objective of (2) at a feasible point λ , thus the objective $h(\lambda)^\top d$ seeks to find a direction d of steepest descent. The first three constraints ensure that d is a feasible direction. The constraint $\|d\|_0 \leq q$ is added to ensure that the direction is of cardinality at most q . Finally, the constraint $\|d\|_\infty \leq 1$ is added to ensure that the program is bounded, and may be viewed as normalizing the scale of the direction d .

The program (9) is, in general, a challenging mixed integer program because of the cardinality constraint. Joachims [1999] discusses an algorithm to solve a similar problem of finding a low cardinality descent direction in an SVM classification setting. Their problem can be easily solved provided that the cardinality q is even, however no such solution seems to exist for our case. However, in our case, when $q = 2$, there is a tractable way to solve (9). We will restrict attention to this special case, i.e., consider only descent directions of cardinality two. In Section A.4.3, we will establish that this is, in fact, sufficient: if the prior iterate λ is suboptimal, then there will exist a direction of descent of cardinality two.

To begin, define the sets

$$\mathcal{P}_1 \triangleq \left\{ (x,a) \in \hat{\mathcal{S}} \times \mathcal{A} \mid \lambda_{a,x} = 0 \right\}, \quad \mathcal{P}_2 = \left\{ x \in \hat{\mathcal{S}} \mid \sum_{a \in \mathcal{A}} \lambda_{x,a} = \frac{\kappa}{N} \right\}.$$

Consider the following procedure:

1. Sort the set of indices $\hat{\mathcal{S}} \times \mathcal{A}$ according to their corresponding component values in the gradient vector $h(\lambda)$. Call this sorted list \mathcal{L}_1 .
2. Denote by (x_1, a_1) the largest element of \mathcal{L}_1 such that $(x_1, a_1) \notin \mathcal{P}_1$, and denote by (x_2, a_2) the smallest element of \mathcal{L}_1 such that $x_2 \notin \mathcal{P}_2$. Add the tuple (x_1, a_1, x_2, a_2) to the list \mathcal{L}_2 .
3. Consider all $x \in \mathcal{P}_2$. For each such x , denote by (x, a_1) the largest element of \mathcal{L}_1 such that $(x, a_1) \notin \mathcal{P}_1$, and denote by (x, a_2) the smallest element of \mathcal{L}_1 . Add each tuple (x, a_1, x, a_2) to \mathcal{L}_2 .
4. Choose the element $(x_1^*, a_1^*, x_2^*, a_2^*)$ of \mathcal{L}_2 that optimizes

$$(10) \quad \min_{(x_1, a_1, x_2, a_2) \in \mathcal{L}_2} h(\lambda)_{x_2, a_2} - h(\lambda)_{x_1, a_1}.$$

Set $d_{x_1^*, a_1^*} = -1$, $d_{x_2^*, a_2^*} = 1$, and all other components of d to zero..

This procedure finds a direction of maximum descent of cardinality two by examining considering candidate index pairs (x_1, a_1, x_2, x_2) for which $h(\lambda)_{x_2, a_1} - h(\lambda)_{x_1, a_1}$ is minimal. Instead of considering at all $N^2 A^2$ such pairs, the routine selectively checks only pairs with describe feasible directions with respect to the constraints of (9). Step 2 considers all pairs with different states, while Step 3 considers all pairs with the same state. It is thus easy to see that the output of this procedure is an optimal solution to (9), i.e., a direction of steepest descent of cardinality two. Further, if the value of the minimal objective (10) determined by this procedure is non-negative, then no descent direction of cardinality two exists, and the algorithm terminates.

In terms of computational complexity, this subset selection step requires us to first compute the gradient of the objective function $h(\lambda) \triangleq Q\lambda + R$. If the gradient is known at the first iteration, then we can update it at each step by generating two columns of Q , since λ only changes at two co-ordinates. Hence, the gradient calculation can be performed in $O(NA)$ time, and with $O(NA)$ storage (since it is not necessary to store Q). For Step 1 of the subset selection procedure, the component indices must be sorted in the order given by the gradient $h(\lambda)$. This operation requires computational effort of the order $O(NA \log NA)$. With the sorted indices, the effort required in the remaining steps to find the steepest direction is via the outlined procedure is $O(NA)$. Thus, our subset selection step requires total computational effort of the order $O(NA \log NA)$.

A .4.2. QP Subproblem

Given a prior iterate $\lambda^{(t-1)}$, and a subset $\mathcal{B} \triangleq \{(x_1, a_1), (x_2, a_2)\}$ of decision variable components of cardinality two as computed in Section A .4.1, we have the restricted optimization problem

$$\begin{aligned}
(11) \quad & \text{minimize} \quad \sum_{(x,a) \in \mathcal{B}} \sum_{(x',a') \in \mathcal{B}} \lambda_{x',a'} Q(x, a, x', a') \lambda_{x,a} \\
& + \sum_{(x,a) \in \mathcal{B}} \lambda_{x,a} \left(R(x, a) + 2 \sum_{(x',a') \notin \mathcal{B}} Q(x, a, x', a') \lambda_{x',a'}^{(t-1)} \right) \\
& \text{subject to} \quad \sum_{a: (x,a) \in \mathcal{B}} \lambda_{x,a} \leq \frac{\kappa}{N} - \sum_{a: (x,a) \notin \mathcal{B}} \lambda_{x,a}^{(t-1)}, \quad \forall x \in \{x_1, x_2\} \\
& \quad \sum_{(x,a) \in \mathcal{B}} \lambda_{x,a} = \frac{1}{1-\alpha} - \sum_{(x,a) \notin \mathcal{B}} \lambda_{x,a}^{(t-1)}, \\
& \quad \lambda \in \mathbb{R}_+^{\mathcal{B}}.
\end{aligned}$$

This subproblem has small dimension. In fact, the equality constraint implies that $\lambda_{x_1, a_1} + \lambda_{x_2, a_2}$ is constant, hence, the problem is in fact a one-dimensional QP that can be solved in closed form. Further, to construct this QP, two columns of Q are required to be generated. This requires computation effort of order $O(NA)$.

Note that the subset \mathcal{B} is chosen so that it is guaranteed to contain a descent direction, according to the procedure in Section A .4.1. Then, the solution of (9) will produce an iterate $\lambda^{(t)}$ that is feasible for the original problem (11) and has lower objective value than the prior iterate $\lambda^{(t-1)}$.

A .4.3. Correctness

The following proposition establishes the correctness of our active set method: if the prior iterate $\lambda \triangleq \lambda^{(t-1)}$ is suboptimal, then there must exist a direction of descent of cardinality two. Our iterative procedure will therefore improve the solution, and will only terminate when global optimality is achieved.

Proposition 2. *If $\lambda \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ is feasible but suboptimal for (2) then, there exists a descent direction of cardinality two.*

The proof of Proposition 2 requires the following lemma:

Lemma 4. Suppose $x, y \in \mathbb{R}^n$ are vectors such that $\mathbf{1}^\top x = 0$ and $x^\top y < 0$. Then there exist co-ordinates $\{i, j\}$, such that $y_i < y_j$, $x_i > 0$, and $x_j < 0$.

Proof. Define the index sets $S^+ \triangleq \{i \mid x_i > 0\}$ and $S^- \triangleq \{i \mid x_i < 0\}$. Under the given hypotheses, both sets are non-empty. Using the fact that $\mathbf{1}^\top x = 0$, define

$$Z \triangleq \sum_{i \in S^+} x_i = \sum_{i \in S^-} (x_i)^-,$$

where $(x_i)^- \triangleq -\min(x_i, 0)$ is the negative part of the scalar x_i . Observe that $Z > 0$. Further, since $x^\top y < 0$,

$$\frac{1}{Z} \sum_{i \in S^+} x_i y_i < \frac{1}{Z} \sum_{i \in S^-} (x_i)^- y_i.$$

Since the weighted average of y over S^- is more than its weighted average over S^+ , we can pick an element in S^+ , i and an element of S^- , j such that $y_i < y_j$. ■

Proof of Proposition 2. If λ is suboptimal, since (2) is a convex quadratic program, there will exist some vector $d \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ is a feasible descent direction at λ . Let $g \triangleq h(\lambda)$ be the gradient at that point. We must have that $g^\top d < 0$, so that it is a descent direction, and that d satisfies the first three constraints of (9), so that it is a feasible direction).

Define

$$\mathcal{T} \triangleq \left\{ x \in \hat{\mathcal{S}} \mid \sum_{a \in \mathcal{A}} \lambda_{x,a} = \frac{\kappa}{N}, \max_{a \in \mathcal{A}} d_{x,a} > 0, \min_{a \in \mathcal{A}} d_{x,a} < 0 \right\}, \quad \mathcal{P}_x \triangleq \{a \in \mathcal{A} \mid d_{x,a} \neq 0\}$$

Consider each $x \in \mathcal{T}$. Without loss of generality, assume that $|\mathcal{P}_x| = 2$, i.e., d has exactly two non-zero components corresponding to the state x . This is justified at the end of the proof. Denote these indices by (x, a_+) and (x, a_-) , so that $d_{x,a_+} > 0$ and $d_{x,a_-} < 0$. From the first constraint of (9), we must have that $d_{x,a_+} \leq (d_{x,a_-})^-$.

There are two cases:

- (i) Suppose $g_{x,a_+} < g_{x,a_-}$, for some $x \in \mathcal{T}$.

Then, we can define a descent direction $\tilde{d} \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ of cardinality two by setting $\tilde{d}_{x,a_+} = 1$, $\tilde{d}_{x,a_-} = -1$, and all other components of \tilde{d} to zero.

- (ii) Suppose that $g_{x,a_+} \geq g_{x,a_-}$, for all $x \in \mathcal{T}$.

For all $x \in \mathcal{T}$, define $\hat{d}_x \triangleq (d_{x,a_-})^- - d_{x,a_+} \geq 0$. Then, the fact that $\sum_{x,a} d_{x,a} = 0$ implies that

$$(12) \quad \sum_{\substack{x \notin \mathcal{T} \\ a \in \mathcal{A}}} d_{x,a} - \sum_{x \in \mathcal{T}} \hat{d}_x = 0.$$

At the same time, for all $x \in \mathcal{T}$, we have that

$$d_{x,a_+} g_{x,a_+} + d_{x,a_-} g_{x,a_-} = -\hat{d}_x g_{x,a_-} + d_{x,a_+} (g_{x,a_+} - g_{x,a_-}) \geq -\hat{d}_x g_{x,a_-}.$$

Then, since d is a descent direction, we have that

$$(13) \quad \sum_{\substack{x \notin \mathcal{T} \\ a \in \mathcal{A}}} d_{x,a} g_{x,a} - \sum_{x \in \mathcal{T}} \hat{d}_x g_{x,a_-} < 0.$$

Now, define the vector $\tilde{d} \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ by

$$\tilde{d}_{x,a} = \begin{cases} d_{x,a} & \text{if } x \notin \mathcal{T}, \\ -\hat{d}_x & \text{if } x \in \mathcal{T} \text{ and } (x,a) = (x,a_-), \\ 0 & \text{otherwise.} \end{cases}$$

Applying Lemma 4 to (12) and (13), there must be a pair of indices (x_1, a_1) and (x_2, a_2) such that $\tilde{d}_{x_1, a_1} > 0$, $\tilde{d}_{x_2, a_2} < 0$ and $g_{x_1, a_1} < g_{x_2, a_2}$. For such (x_1, a_1) and (x_2, a_2) we have a descent direction where we can increase λ_{x_1, a_1} and decrease λ_{x_2, a_2} by the same amount and get a decrease in the objective. Note that since $\tilde{d}_{x_1, a_1} > 0$, we have that $d_{x_1, a_1} > 0$ and $x_1 \notin \mathcal{T}$, hence $\sum_a \lambda_{x_1, a} < \kappa/N$. Also, by construction, (x_2, a_2) is chosen such that $d_{x_2, a_2} < 0$, implying that $\lambda_{x_2, a_2} > 0$. Thus the specified direction is also feasible, and we have a feasible descent direction of cardinality two.

Finally, to complete the proof, consider the case where there are some $x \in \mathcal{T}$ with $|\mathcal{P}_x| > 2$, i.e., d has more than two non-zero components corresponding to the state x . For each $x \in \mathcal{T}$, define

$$\begin{aligned} \mathcal{A}_x^+ &\triangleq \{a \in \mathcal{A} \mid d_{x,a} > 0\}, & \mathcal{A}_x^- &\triangleq \{a \in \mathcal{A} \mid d_{x,a} < 0\}, \\ a_1 &\in \operatorname{argmin}_{a \in \mathcal{A}_x^+} g_{x,a}, & a_2 &\in \operatorname{argmax}_{a \in \mathcal{A}_x^-} g_{x,a}. \end{aligned}$$

Define a new direction $\tilde{d} \in \mathbb{R}^{\hat{\mathcal{S}} \times \mathcal{A}}$ by

$$\tilde{d}_{x,a} = \begin{cases} \sum_{a' \in \mathcal{A}_x^+} d_{x,a'} & \text{if } x \in \mathcal{T} \text{ and } (x,a) = (x,a_1), \\ \sum_{a' \in \mathcal{A}_x^-} d_{x,a'} & \text{if } x \in \mathcal{T} \text{ and } (x,a) = (x,a_2), \\ d_{x,a} & \text{otherwise.} \end{cases}$$

It is easy to verify that \tilde{d} is also a feasible descent direction. Furthermore, \tilde{d} has only two non-zero components corresponding to each start $x \in \mathcal{T}$. ■

References

- P. L. Bartlett and S. Mendelson. Journal of machine learning research. *Rademacher and Gaussian Complexities: Risk Bounds and Structural Results*, 3:463–482, 2002.
- V. V. Desai, V. F. Farias, and C. C. Moallemi. Approximate dynamic programming via a smoothed linear program. *Operations Research*, 2011.
- T. Joachims. Making large scale svm learning practical. 1999.
- D.G Luenberger. *Optimization by vector space methods*. John Wiley and Sons, 1997.
- E. Osuna, R. Freund, and F. Girosi. An improved training algorithm for support vector machines. In *Neural Networks for Signal Processing [1997] VII. Proceedings of the 1997 IEEE Workshop*, pages 276–285, sep 1997. doi: 10.1109/NNSP.1997.622408.