A ALTERNATIVE RKHS POLICY PARAMETERIZATION

In Bagnell (2004) the policy

$$\pi_{h,\mathbf{\Sigma}}(a|s) := \frac{1}{Z_s} e^{f(s,a)}$$
$$Z_s = \int_{\mathcal{A}} e^{f(s,a)} \mathrm{d}a$$

where $f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is an element of an RKHS on $\mathcal{S} \times \mathcal{A}$, is considered[6], as in (20). The gradient is computed as follows

$$\begin{aligned}
\frac{\nabla_f \pi_f(a|s)}{\pi_f(a|s)} &= \nabla_f \log \pi_f(a|s) \\
&= \nabla_f \left(f(a|s) - \log \int_{\mathcal{A}} e^{f(a',s)} \mathrm{d}a' \right) \\
&= K((a,s),\cdot) - \frac{\nabla_f \int_{\mathcal{A}} e^{f(a',s)} \mathrm{d}a'}{\int_{\mathcal{A}} e^{f(a',s)} \mathrm{d}a'} \\
&= K((a,s),\cdot) - \frac{\int_{\mathcal{A}} \nabla_f e^{f(a',s)} \mathrm{d}a'}{Z_s} \\
&= K((a,s),\cdot) - \frac{1}{Z_s} \int_{\mathcal{A}} K((a',s),\cdot) e^{f(a',s)} \mathrm{d}a' \\
&= K((a,s),\cdot) - \mathbb{E}_{A \sim \pi_f(\cdot|s)}[K((A,s),\cdot)].
\end{aligned}$$

B GREEDY FEATURE SELECTION USING VECTOR VALUED MATCHING PURSUIT

Our algorithm uses, as a subcomponent, a vector-valued version of the matching pursuit algorithm (Mallat and Zhang, 1993). Although this is a straightforward extension of the scalar case, it is to our knowledge not explicitly derived in the literature and so we derive the method here for clarity.

Suppose we wish to regress a vector-valued function

$$f^* : \mathcal{X} \to \mathcal{V},$$

given a data sample $\mathcal{D} = \{x_i, v_i\}_{i=1}^m$ where $v_i = f^*(x_i) + \epsilon$ where ϵ is zero-mean noise, $f^*(x_i) = \mathbb{E}[V_i|x_i]$. Suppose we are given a *dictionary* $\mathcal{G} = \{g_1, ..., g_n\}$, where $g_i : \mathcal{X} \to \mathbb{R}$, of candidate real-valued functions, and we aim to find an estimate $\hat{f}$ for f^* of the form,

$$\hat{f} = \sum_{i=1}^{D} w^i \hat{g}_i$$

where $\mathcal{B}_D = \{\hat{g}_i\}_{i=1}^D \subseteq \mathcal{G}$ is called the basis and $w^i \in \mathcal{V}$. When $\mathcal{V} = \mathbb{R}$, matching pursuit Mallat and Zhang (1993) can be used to incrementally build the basis, and we now detail the extension to the vector-valued output case. We build the basis incrementally and for each basis $\mathcal{B}_j$ we form an estimate $\hat{f}^j = \sum_{i=1}^j w^i \hat{g}_i$. We begin with the empty basis $\mathcal{B}_0$ and add new basis elements $\hat{g}_{j+1}$ to greedily optimize the objective. For each estimate we define the residue $\boldsymbol{r}^j$,

$$r_i^j = v_i - \hat{f}_j(x_i) \in \mathcal{V},$$

and pick the $g \in \mathcal{D}$ which minimizes the next residue when added to the current estimate,

$$\begin{aligned}
g_{j+1} &= \operatorname*{argmin}_{g \in \mathcal{D}} \min_{w \in \mathcal{V}} \sum_{i=1}^m ||v_i - ((\hat{f}_j + wg)(x_i))||_{\mathcal{V}}^2 \\
&= \operatorname*{argmin}_{g \in \mathcal{D}} \min_{w \in \mathcal{V}} \sum_{i=1}^m ||r_i^j - wg(x_i)||_{\mathcal{V}}^2.
\end{aligned}$$

Since $\nabla_w \sum_{i=1}^m ||r_i^j - wg(x_i)||_{\mathcal{V}}^2 = 0$ at the minimum we have,

$$\begin{aligned}
0 &= \sum_{i=1}^m \nabla_w \left(\langle g(x_i)w, g(x_i)w \rangle_{\mathcal{V}} - 2\langle g(x_i)w, r_i^j \rangle_{\mathcal{V}} \right) \\
&= \sum_{i=1}^m 2wg(x_i)^2 - 2g(x_i)r_i^j \\
w^{j+1} &= \left(\sum_{i=1}^m g(x_i) r_i^j \right) / \left(\sum_{i=1}^m g(x_i)^2 \right) \in \mathcal{V}
\end{aligned}$$

Then,

$$\begin{aligned}
&\sum_{i=1}^m ||r_i^j - w^{j+1} g(x_i)||_{\mathcal{V}}^2 \\
&= \sum_{i=1}^m ||r_i^j||_{\mathcal{V}}^2 - 2\sum_{i=1}^m g(x_i) \langle r_i^j, w^{\min} \rangle_{\mathcal{V}} + ||w^{\min}||_{\mathcal{V}}^2 \sum_{i=1}^m g(x_i)^2 \\
&= \sum_{i=1}^m ||r_i^j||_{\mathcal{V}}^2 - \frac{2\sum_{i=1}^m g(x_i) \langle r_i^j, \sum_{k=1}^m g(x_k) r_k^j \rangle_{\mathcal{V}}}{\sum_{k=1}^m g(x_k)^2} \\
&\quad + \frac{||\sum_{k=1}^m g(x_k) r_k^j||_{\mathcal{V}}^2}{\sum_{i=1}^m g(x_i)^2} \\
&= \sum_{i=1}^m ||r_i^j||_{\mathcal{V}}^2 - \frac{||\sum_{i=1}^m g(x_i) r_i^j||_{\mathcal{V}}^2}{\sum_{i=1}^m g(x_i)^2}
\end{aligned} \tag{23}$$

Thus $\hat{g}_{j+1} = \operatorname{argmax}_{g \in \mathcal{G}} \frac{||\sum_{i=1}^m g(x_i) r_i^j||_{\mathcal{V}}^2}{\sum_{i=1}^m g(x_i)^2}$. Thus at each iteration of matching pursuit we must evaluate $\frac{||\sum_{i=1}^m g(x_i) r_i^j||_{\mathcal{V}}^2}{\sum_{i=1}^m g(x_i)^2}$ for a selection of k dictionary elements (not necessarily all). We have,

$$||\sum_{i=1}^m g(x_i) r_i^j||_{\mathcal{V}}^2 = ||\sum_{i=1}^m g(x_i)(\hat{f}^j(x_i) - v_i)||_{\mathcal{V}}^2$$

[6] strictly speaking the function f is defined on observations rather than states in Bagnell (2004)

For each dictionary element g this can be computed in $O(mj + md + jd)$ where $d = \dim(\mathcal{V})$, and so $O(k(mj + md + jd))$ over k dictionary elements.

It is sometimes useful, at iteration j to "backfit" all the weights $\{w^i\}_{i=1}^j$ by replacing them with the least squares solution: i.e. matching pursuit is used to find the basis but the weights are finally optimized using least squares. Alternatively this can be performed end of the process or several times throughout.

In order to find a compact representation we can also use matching pursuit adaptively by setting a tolerance δ such that the algorithm terminates when it fails to reduce the residue by more than δ. Thus the method will only add features if they significantly help.

The output of vector valued matching pursuit is a collection of weights $\{w^i\}_{i=1}^j$ and features $\{\hat{g}_i(\cdot)\}_{i=1}^j$ such that $f^* \approx \sum_{i=1}^j w^i \hat{g}_i$.