Natural Policy Gradient for Exponential Families

Carson Eisenach* Zhuoran Yang*

February 18, 2019

Abstract

Recent work has highlighted how a misalignment between the support of the policy and the action space of the reinforcement learning problem can introduce bias and unnecessary variance into policy gradient estimates. To better align the support of the policy and the action space, we can consider using arbitrary exponential families to model the policy distribution. Exponential families are a natural choice because the class of exponential families is very rich and can model the support of most action spaces of practical interest. While the multivariate Gaussian is the most commonly used distribution today, in general it is possible to efficiently implement both natural policy gradient and TRPO for any exponential family. In this technical report we derive efficient natural policy gradient update rules for several exponential familes. We also consider an application of the Gamma distribution to an optimal production problem and show that it substantially outperforms the Gaussian.

1 Introduction

Deep reinforcement learning has seen great success in a wide variety of application areas – ranging from complex games like Go (Silver et al., 2016) to high dimensional continuous control (Schulman et al., 2015, 2016). For problems with discrete controls, such as Go, value based methods have proved invaluable. For continuous control, is is more natural to model the policy directly, using so-called policy based methods. However, it is widely acknowledged that policy based methods such as Policy Gradient and TRPO are not nearly as robust as value based methods due to the difficulty of obtaining low-variance unbiased estimators. Recent work (Chou et al., 2017; Fujita and Maeda, 2018; Eisenach et al., 2018) has demonstrated that part of the difficulty stems from a mismatch of support in the *sampling policy* to the support of the *effective action space* – indeed the authors show how bias and unnecessary variance is introduced into the learning process, and propose methods to counteract it.

In this work, we derive algorithms tailored to *exponential family* policy distributions. Exponential families are a natural choice because (1) they have many desirable properties including easy to derive update rules for a wide range of algorithms and (2) exponential families are very rich, and in

*Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08540.

all practical cases, one can find an exponential family with appropriate support for the problem of interest. Because it is easy to derive quantities such as the Fisher Information for exponential families, one can obtain significantly more efficient implementations of popular algorithms like TRPO and natural policy gradient. We validate the use of exponential families experimentally on several benchmark continuous control problems. By moving beyond Gaussian and Beta distributions to arbitrary exponential families, we can tailor the policy distribution to the action space of a particular RL problem.

1.1 Natural Policy Gradient

The natural policy gradient update is given by

$$\widetilde{\nabla} J(\theta) = F^{-1}(\theta)\nabla J(\theta),$$

where ∇J is the standard policy gradient, $F(\theta) = \mathbb{E}_{\rho(\theta),\pi(\theta)}\left[\nabla \log f_\pi(a|s)\nabla \log f_\pi(a|s)^\top\right]$, and $\rho(\theta)$ is the discounted state occupancy distribution (or invariant distribution if we work in the average reward setting).

1.2 Exponential Family

Members of the exponential family have densities of the form

$$f(x;\eta) = h(x)\exp\left[T(x)^\top \eta - A(\eta)\right]$$

where T is the sufficient statistic, η the natural parameter, and A is the log partition function.

If our policy is a member of the exponential family, and we parametrize the natural parameter η by $\theta \in \Theta$, then the score function at state s becomes

$$\mathrm{D}_\theta \log f_\pi(a) = T(a)^\top \mathrm{D}_\theta \eta(\theta) - \mathrm{D}_\eta A(\eta)\mathrm{D}_\theta \eta(\theta),$$

where for now we suppress the dependence on the state s. To simplify the expression above, we can use the fact that $\mathrm{D}_\eta A(\eta) = \mathbb{E}_\eta\left[T(a)\right]^\top$ and by plugging into the display above get that

$$\mathrm{D}_\theta \log f_\pi(a) = \left[T(a) - \mathbb{E}_\eta\left[T(a)\right]\right]^\top \mathrm{D}_\theta \eta(\theta). \tag{1.1}$$

To obtain the natural gradient when η is a function of θ, we first observe that the fisher information is given by

$$\mathcal{I}(\theta) = \mathrm{D}_\theta \eta^\top \mathcal{I}(\eta)\mathrm{D}_\theta \eta. \tag{1.2}$$

More generally, we can observe that (1.2) holds for any change of parametrization. We will leverage this result in the sequel.

1.3 Why Use Exponential Families?

There are several reasons we are interested in using exponential families. One reason is that the Fisher information, with respect to the natural parameters η, is given by

$$\mathcal{I}(\eta) = \mathrm{D}_\eta \mathbb{E}_\eta\left[T(x)\right]^\top. \tag{1.3}$$

2 Derivations for Gaussian Families

2.1 Single Variate Gaussian

We mention, without derivation, that the score function for the single-variate Gaussian is given by

$$\psi(x;\omega) = \left[\sigma^{-2}(x-\mu); -\frac{1}{2}\sigma^{-2} + \frac{1}{2}\sigma^{-4}(x-\mu)^2\right]^\top, \tag{2.1}$$

where $\omega = (\mu, \sigma^2)$ is the moment parameter.

2.2 Multivariate Gaussian

Recall that the density of the multivariate Gaussian in terms of its moment parameters μ and Σ is given by

$$p(x;\mu,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\Big((x-\mu)^\top\Sigma^{-1}(x-\mu)\Big)$$

where $x \in \mathbb{R}^d$. As a member of the exponential family, we can find that

$$T(x) = \left[x, xx^\top\right], \eta = \left[\Sigma^{-1}\mu; -\frac{1}{2}\Sigma^{-1}\right], A(\eta) = \frac{1}{2}\left(\mu^\top\Sigma^{-1}\mu + \log|\Sigma|\right), h(x) = (2\pi)^{-d/2}.$$

Later we will find it useful to sometimes work instead with the vectorized parameter $\widetilde{\eta} := [\eta_1; \mathrm{vec}(\eta_2)]$.

Jacobian of natural parameters with respect to the moment parameters

The moment parameters are given by $\omega = (\mu, \mathrm{vec}(\Sigma))$. To obtain the Jacobian of the transformation, we first observe that $\mathrm{d}\eta_1 = -\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}\mu + \Sigma^{-1}\mathrm{d}\mu$, and therefore

$$\mathrm{D}_\omega\widetilde{\eta}_1 = \left[\Sigma^{-1}; -\mu^\top\Sigma^{-1}\otimes\Sigma^{-1}\right].$$

Likewise, $\mathrm{d}\eta_2 = \frac{1}{2}\Sigma^{-1}(\mathrm{d}\Sigma)\Sigma^{-1}$, or equivalently that

$$\mathrm{D}_\omega\widetilde{\eta}_2 = \left[0; \frac{1}{2}\Sigma^{-1}\otimes\Sigma^{-1}\right].$$

Combining the two previous displays gives

$$\mathrm{D}_\omega\widetilde{\eta} = \begin{bmatrix}\Sigma^{-1} & -\mu^\top\Sigma^{-1}\otimes\Sigma^{-1}\\ 0 & \frac{1}{2}\Sigma^{-1}\otimes\Sigma^{-1}\end{bmatrix}. \tag{2.2}$$

Score Function

Recall the score function is defined as $\psi(x;\omega) = \mathrm{D}f(x;\omega)^\top$. Using (1.1) and (2.2), we can obtain that

$$\psi_1(x;\omega) = \Sigma^{-1}(x-\mu)$$

and

$$\begin{aligned}
\psi_2(x;\omega) &= \left[(x-\mu)^\top\left(-\mu^\top\Sigma^{-1}\otimes\Sigma^{-1}\right)+\mathrm{vec}\left(xx^\top-\Sigma^{-1}-\mu\mu^\top\right)^\top\left(\frac{1}{2}\Sigma^{-1}\otimes\Sigma^{-1}\right)\right]^\top \\
&= -\left(\Sigma^{-1}\mu\otimes\Sigma^{-1}\right)(x-\mu)+\frac{1}{2}\left(\Sigma^{-1}\otimes\Sigma^{-1}\right)\mathrm{vec}\left(xx^\top-\Sigma^{-1}-\mu\mu^\top\right) \\
&= -\mathrm{vec}\left(\Sigma^{-1}\left(x\mu^\top-\mu\mu^\top\right)\Sigma^{-1}\right)+\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\left(xx^\top-\Sigma-\mu\mu^\top\right)\Sigma^{-1}\right) \\
&= \mathrm{vec}\left(\Sigma^{-1}\left(\frac{1}{2}xx^\top-x\mu^\top+\frac{1}{2}\mu\mu^\top-\frac{1}{2}\Sigma\right)\Sigma^{-1}\right) \\
&= \frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}(x-\mu)(x-\mu)^\top\Sigma^{-1}-\Sigma^{-1}\right)
\end{aligned}$$

where we used $\mathbb{E}\left[xx^\top\right]=\Sigma+\mu\mu^\top$. Thus in terms of the moment parameters $\omega=(\mu,\mathrm{vec}(\Sigma))$ the score function is given by

$$\psi(x;\omega)=\left[(x-\mu)^\top\Sigma^{-1};\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}(x-\mu)(x-\mu)^\top\Sigma^{-1}-\Sigma^{-1}\right)^\top\right]^\top, \tag{2.3}$$

which in 1-dimension clearly reduces to (2.1).

Fisher Information Matrix - Derivation 1

In this first approach, we take the Fisher Information with respect to the canonical parameters, and then transfer to the moment parametrization using (1.2). As in (1.1), we see that for any exponential family

$$\mathrm{d}l(x,\eta)=\sum_i\left[T_i(x)-\mathbb{E}\left[T_i(x)\right]\right]^\top\mathrm{d}\eta_i$$

where $l(\eta)$ is the log-likelihood function; which inner product is denoted by $x^\top x$ should be clear from the context. For the Gaussian, this becomes

$$\mathrm{d}l(x,\eta)=\left[x+\frac{1}{2}\eta_2^{-1}\eta_1\right]^\top\mathrm{d}\eta_1+\mathrm{tr}\left[\left(xx^\top+\frac{1}{2}\eta_2^{-1}-\frac{1}{4}\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\right)\mathrm{d}\eta_2\right], \tag{2.4}$$

where we used the identities $\mu=-\frac{1}{2}\eta_2^{-1}\eta_1$ and $\Sigma=-\frac{1}{2}\eta_2^{-1}$. Taking the differential a second time,

$$\mathrm{d}^2l(x,\eta)=\underbrace{\mathrm{d}\left[x+\frac{1}{2}\eta_2^{-1}\eta_1\right]^\top\mathrm{d}\eta_1}_{\text{(i)}}+\underbrace{\mathrm{d}\,\mathrm{tr}\left[\left(xx^\top+\frac{1}{2}\eta_2^{-1}-\frac{1}{4}\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\right)\mathrm{d}\eta_2\right]}_{\text{(ii)}}. \tag{2.5}$$

Immediately, term (i) can be expanded as

$$\begin{aligned}
\text{(i)} &= \frac{1}{2}\mathrm{d}\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_1-\frac{1}{2}\eta_1^\top\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\mathrm{d}\eta_1 \\
&= \frac{1}{2}\mathrm{d}\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_1-\frac{1}{2}\left(\mathrm{dvec}\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\otimes\eta_2^{-1}\right)\mathrm{d}\eta_1.
\end{aligned}$$

For term (ii), we see that

$$
\begin{aligned}
\text{(ii)} &= -\frac{1}{2}\operatorname{tr}\left(\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\mathrm{d}\eta_2\right) - \frac{1}{4}\operatorname{tr}\left(\left[\mathrm{d}\left(\eta_2^{-1}\right)\eta_1\eta_1^\top\eta_2^{-1} + \eta_2^{-1}\mathrm{d}\left(\eta_1\eta_1^\top\right)\eta_2^{-1} + \eta_2^{-1}\eta_1\eta_1^\top\mathrm{d}\left(\eta_2^{-1}\right)\right]\mathrm{d}\eta_2\right) \\
&= -\frac{1}{2}\operatorname{tr}\left(\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\mathrm{d}\eta_2\right) - \frac{1}{4}\underbrace{\operatorname{tr}\left(\mathrm{d}\left(\eta_2^{-1}\right)\eta_1\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_2 + \eta_2^{-1}\eta_1\eta_1^\top\mathrm{d}\left(\eta_2^{-1}\right)\mathrm{d}\eta_2\right)}_{\text{(ii.a)}} \\
&\quad - \frac{1}{4}\underbrace{\operatorname{tr}\left(\eta_2^{-1}\mathrm{d}\left(\eta_1\eta_1^\top\right)\eta_2^{-1}\mathrm{d}\eta_2\right)}_{\text{(ii.b)}}
\end{aligned} \tag{2.6}
$$

We can expand term (ii.a) as

$$
\begin{aligned}
\text{(ii.a)} &= -\operatorname{tr}\left(\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_2\right) - \operatorname{tr}\left(\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\mathrm{d}\eta_2\right) \\
&= -2\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_2\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\eta_1 \\
&= -2\left(\mathrm{d}\eta_2\eta_2^{-1}\eta_1\right)^\top\left(\eta_2^{-1}\left(\mathrm{d}\eta_2\right)\eta_2^{-1}\eta_1\right) \\
&= -2\left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_1^\top\eta_2^{-1}\otimes I_{d,d}\right)^\top\left(\eta_1^\top\eta_2^{-1}\otimes\eta_2^{-1}\right)\mathrm{dvec}\,\eta_2 \\
&= -2\left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\otimes\eta_2^{-1}\right)\mathrm{dvec}\,\eta_2.
\end{aligned}
$$

Likewise term (ii.b) can be expanded as

$$
\begin{aligned}
\text{(ii.b)} &= \operatorname{tr}\left(\eta_2^{-1}\left(\mathrm{d}\eta_1\right)\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_2\right) + \operatorname{tr}\left(\eta_2^{-1}\eta_1\left(\mathrm{d}\eta_1\right)^\top\eta_2^{-1}\mathrm{d}\eta_2\right) \\
&= \eta_1^\top\eta_2^{-1}\mathrm{d}\eta_2\eta_2^{-1}\mathrm{d}\eta_1 + \left(\mathrm{d}\eta_1\right)^\top\eta_2^{-1}\mathrm{d}\eta_2\eta_2^{-1}\eta_1 \\
&= 2\left(\eta_2^{-1}\mathrm{d}\eta_2\eta_2^{-1}\eta_1\right)^\top\mathrm{d}\eta_1 \\
&= 2\left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\otimes\eta_2^{-1}\right)\mathrm{d}\eta_1
\end{aligned}
$$

Plugging back into (2.6) gives that

$$
\text{(ii)} = -\frac{1}{2}\left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\otimes\eta_2^{-1}\right)\mathrm{d}\eta_1 + \frac{1}{2}\left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\otimes\eta_2^{-1} - \eta_2^{-1}\otimes\eta_2^{-1}\right)\mathrm{dvec}\,\eta_2
$$

Plugging the expressions for (i) and (ii) into (2.5) gives

$$
\begin{aligned}
\mathrm{d}^2 l(x,\eta) &= \frac{1}{2}\mathrm{d}\eta_1^\top\eta_2^{-1}\mathrm{d}\eta_1 - \left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\otimes\eta_2^{-1}\right)\mathrm{d}\eta_1 \\
&\quad + \frac{1}{2}\left(\mathrm{dvec}\,\eta_2\right)^\top\left(\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\otimes\eta_2^{-1} - \eta_2^{-1}\otimes\eta_2^{-1}\right)\mathrm{dvec}\,\eta_2
\end{aligned} \tag{2.7}
$$

Applying Lemma A.2 to (2.7) gives

$$
\mathrm{H}_{\widetilde{\eta}}\, l(x,\widetilde{\eta}) = \frac{1}{2}\begin{bmatrix} \eta_2^{-1} & -\eta_1^\top\eta_2^{-1}\otimes\eta_2^{-1} \\ -\eta_2^{-1}\eta_1\otimes\eta_2^{-1} & \left(\eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1} - \eta_2^{-1}\right)\otimes\eta_2^{-1} \end{bmatrix},
$$

and therefore that

$$
\mathcal{I}(\widetilde{\eta}) = \frac{1}{2}\begin{bmatrix} -\eta_2^{-1} & \eta_1^\top\eta_2^{-1}\otimes\eta_2^{-1} \\ \eta_2^{-1}\eta_1\otimes\eta_2^{-1} & \left(\eta_2^{-1} - \eta_2^{-1}\eta_1\eta_1^\top\eta_2^{-1}\right)\otimes\eta_2^{-1} \end{bmatrix} = \begin{bmatrix} \Sigma & 2\mu^\top\otimes\Sigma \\ 2\mu\otimes\Sigma & \left(2\Sigma + 4\mu\mu^\top\right)\otimes\Sigma \end{bmatrix}. \tag{2.8}
$$

To transfer to the moment parametrization, we first compute the intermediate result $A := \mathrm{D}_\omega \eta^\top \mathcal{I}(\eta)$ by using (2.8) and (2.2). This gives us

$$A = \begin{bmatrix} I_{d,d} & 2\mu^\top \otimes I_{d,d} \\ 0_{d^2,d} & I_{d,d} \otimes I_{d,d} \end{bmatrix}.$$

Finally, by multiplying again by $\mathrm{D}_\omega \widetilde{\eta}$, we obtain

$$\mathcal{I}(\omega) = \begin{bmatrix} \Sigma^{-1} & 0 \\ 0 & \frac{1}{2}\Sigma^{-1} \otimes \Sigma^{-1} \end{bmatrix}, \tag{2.9}$$

which as we can see, matches the Fisher information for the single-variate Gaussian with respect to the moment parameters.

2.3 Multivariate Gaussian with Diagonal Covariance

In practice, we often model the covariance as diagonal (ie. $\Sigma = \mathrm{diag}(\sigma)$), in which case the previous display can be simplified to

$$T(x) = \left[x; x^2\right]^\top, \eta = \left[\sigma^{-1} \circ \mu; -\frac{1}{2}\sigma^{-1}\right], A(\eta) = \frac{1}{2}\left(\mu^\top \mathrm{diag}(\sigma^{-1})\mu + \log|\mathrm{diag}(\sigma)|\right), h(x) = (2\pi)^{-d/2},$$

where all exponentials are to be interpreted as entry-wise.

Jacobian of natural parameters with respect to the moment parameters

The moment parameters are given by $\omega = (\mu, \sigma)$. To obtain the Jacobian of the transformation, we first observe that $\mathrm{d}\eta_1 = -\sigma^{-2} \circ \mu \circ \mathrm{d}\sigma + \sigma^{-1} \circ \mathrm{d}\mu$, and therefore

$$\mathrm{D}_\omega \eta_1 = \left[\mathrm{diag}(\sigma^{-1}); -\mathrm{diag}(\sigma^{-2} \circ \mu)\right].$$

Likewise, $\mathrm{d}\eta_2 = \frac{1}{2}\sigma^{-2} \circ \mathrm{d}\sigma$, or equivalently that

$$\mathrm{D}_\omega \eta_2 = \left[0; \frac{1}{2}\mathrm{diag}(\sigma^{-2})\right].$$

Combining the two previous displays gives

$$\mathrm{D}_\omega \widetilde{\eta} = \begin{bmatrix} \mathrm{diag}(\sigma^{-1}) & -\mathrm{diag}(\sigma^{-2} \circ \mu) \\ 0 & \frac{1}{2}\mathrm{diag}(\sigma^{-2}) \end{bmatrix}. \tag{2.10}$$

Score Function

If we view η as a function of $\omega := (\mu, \sigma)$, then we can reduce (2.3) to

$$\psi(x, \omega) = \left[\left((x - \mu) \circ \sigma^{-1}\right)^\top; \frac{1}{2}\left(\left(x^2 - \mu^2\right) \circ \sigma^{-2} - \sigma^{-1}\right)^\top\right]^\top. \tag{2.11}$$

Fisher Information

Though we could derive the appropriate expressions for this model from the previous section, it will be just as easy to re-derive the results directly. We can find the first differential of the log-likelihood as

$$\mathrm{d}l(x,\eta) = \left[x + \frac{1}{2}\eta_2^{-1} \circ \eta_1\right]^\top \mathrm{d}\eta_1 + \left[xx^\top + \frac{1}{2}\eta_2^{-1} - \frac{1}{4}\eta_2^{-2} \circ \eta_1^2\right]^\top \mathrm{d}\eta_2. \tag{2.12}$$

Taking the differential a second time, we see that

$$\begin{aligned}\mathrm{d}^2 l(x,\eta) &= \frac{1}{2}\left[\eta_2^{-1} \circ \mathrm{d}\eta_1 - \eta_1 \circ \eta_2^{-2} \circ \mathrm{d}\eta_2\right]^\top \mathrm{d}\eta_1 + \frac{1}{2}\left[-\eta_2^{-2} \circ \mathrm{d}\eta_2 + \eta_2^{-3} \circ \eta_1^2 \circ \mathrm{d}\eta_2 - \eta_2^{-2} \circ \eta_1 \circ \mathrm{d}\eta_1\right]^\top \mathrm{d}\eta_2 \\ &= \frac{1}{2}\mathrm{d}\eta_1^\top \mathrm{diag}(\eta_2^{-1})\mathrm{d}\eta_1 - \mathrm{d}\eta_2^\top \mathrm{diag}(\eta_1 \circ \eta_2^{-2})\mathrm{d}\eta_1 + \frac{1}{2}\mathrm{d}\eta_2^\top \mathrm{diag}(\eta_2^{-3} \circ \eta_1^2 - \eta_2^{-2})\mathrm{d}\eta_2. \end{aligned} \tag{2.13}$$

Applying Lemma A.2 to (2.13) gives

$$\mathrm{H}_\eta l(x,\eta) = \frac{1}{2}\begin{bmatrix} \eta_2^{-1} & -\mathrm{diag}(\eta_1 \circ \eta_2^{-2}) \\ -\mathrm{diag}(\eta_1 \circ \eta_2^{-2}) & \mathrm{diag}(\eta_2^{-3} \circ \eta_1^2 - \eta_2^{-2}) \end{bmatrix},$$

and therefore that

$$\mathcal{I}(\eta) = \frac{1}{2}\begin{bmatrix} -\mathrm{diag}(\eta_2^{-1}) & \mathrm{diag}(\eta_1 \circ \eta_2^{-2}) \\ \mathrm{diag}(\eta_1 \circ \eta_2^{-2}) & \mathrm{diag}(\eta_2^{-2} - \eta_2^{-3} \circ \eta_1^2) \end{bmatrix} = \begin{bmatrix} \mathrm{diag}(\sigma) & 2\mathrm{diag}(\sigma^2 \circ \mu) \\ 2\mathrm{diag}(\sigma^2 \circ \mu) & \mathrm{diag}(2\sigma^2 + 4\sigma \circ \mu^2) \end{bmatrix}. \tag{2.14}$$

To transfer to the moment parameters, we use (1.2), which gives us

$$\mathcal{I}(\omega) = \begin{bmatrix} \mathrm{diag}(\sigma^{-1}) & 0 \\ 0 & \frac{1}{2}\mathrm{diag}(\sigma^{-2}) \end{bmatrix}. \tag{2.15}$$

If instead we choose to parametrize by $\omega' := (\mu, \sigma^{1/2})$, the mean and standard deviation, then using the fact that $\mathrm{D}_{\sigma^{1/2}}\sigma = 2\sigma^{1/2}$ and (1.2) we find that

$$\mathcal{I}(\omega') = \begin{bmatrix} \mathrm{diag}(\sigma^{-1}) & 0 \\ 0 & 2\mathrm{diag}(\sigma^{-1}) \end{bmatrix}. \tag{2.16}$$

A final parametrization of interest is $\omega'' := (\mu, \frac{1}{2}\log\sigma)$. In this instance, similarly find that $\mathrm{D}_{1/2\log\sigma}\sigma = 2\sigma^{1/2}$

$$\mathcal{I}(\omega'') = \begin{bmatrix} \mathrm{diag}(\sigma^{-1}) & 0 \\ 0 & 2 \end{bmatrix}. \tag{2.17}$$

Modeling the Moment Parameters

We model σ directly and μ as $\mu = f(\theta_\mu)$, giving the parameter $\theta := (\theta_\mu, \sigma)$. Then the Fisher information is given by

$$\mathcal{I}(\theta) = \mathrm{D}_\theta\omega(\theta)^\top \mathcal{I}(\omega)\mathrm{D}_\theta\omega(\theta),$$

where

$$\mathrm{D}_\theta\omega(\theta) = \begin{bmatrix} \mathrm{D}_\theta f & 0 \\ 0 & \mathbf{I} \end{bmatrix}.$$

Combining these expressions and multiplying out gives

$$\mathcal{I}(\theta) = \begin{bmatrix} \mathrm{D}_\theta f^\top \mathcal{I}(\omega)_{1,1}\mathrm{D}_\theta f & \mathrm{D}_\theta f^\top \mathcal{I}(\omega)_{1,2} \\ \mathcal{I}(\omega)_{1,2}\mathrm{D}_\theta f & \mathcal{I}(\omega)_{2,2} \end{bmatrix}. \tag{2.18}$$

3 Multivariate Beta Distribution

If a random variable $X \in [0,1]$ is distributed according to the beta distribution with parameters (α, β), then it has the density

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}.$$

We consider the random vector $x \in \mathbb{R}^d$ distributed according to the product d independent beta distributions, each of which is parametrized by some α_i, β_i. In particular,

$$f(x;\alpha,\beta) = \prod_{i=1}^{d} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x_i^{\alpha_i-1}(1-x_i)^{\beta_i-1}.$$

The log-likelihood function is given by

$$l(\alpha,\beta;x) = \sum_i (\alpha_i - 1)\log(x_i) + (\beta_i - 1)\log(1-x_i) + \log\Gamma(\alpha_i+\beta_i) - \log\Gamma(\alpha_i) - \log\Gamma(\beta_i).$$

Useful Properties

It will often be useful to write results in terms of the *polygamma* functions, defined as

$$\psi^{(m)}(x) := \frac{d^{m+1}}{dx^{m+1}} \log\Gamma(x).$$

One fact that will be useful later is that for the single variate beta-distribution, $\mathbb{E}[\log x] = \psi^0(\alpha) - \psi^0(\alpha+\beta)$ and $\mathbb{E}[\log(1-x)] = \psi^0(\beta) - \psi^0(\alpha+\beta)$.

3.1 Natural Parametrization

In this section we show that in fact $\eta = [\alpha; \beta]$ is the natural parameter of the multi-variate beta distribution. Indeed,

$$\begin{aligned}
f(x;\alpha,\beta) &= \left[\prod_{i=1}^{d} \frac{\Gamma(\alpha_i+\beta_i)}{\Gamma(\alpha_i)\Gamma(\beta_i)} x_i^{\alpha_i}(1-x_i)^{\beta_i}\right]\left[\prod_{i=1}^{d} x_i^{-1}(1-x_i)^{-1}\right] \\
&= h(x)\exp\left[\sum_{i=1}^{d} \alpha_i \log(x_i) + \beta_i \log(1-x_i) + \log\Gamma(\alpha_i+\beta_i) - \log\Gamma(\alpha_i) - \log\Gamma(\beta_i)\right] \\
&= h(x)\exp\left[T(x)^\top \eta - A(\eta)\right]
\end{aligned}$$

where $h(x) = \left[\prod_{i=1}^{d} x_i^{-1}(1-x_i)^{-1}\right]$, $T(x) = [\log(x); \log(1-x)]$, and $A(\eta) = \sum_{i=1}^{d} \log\frac{\Gamma(\alpha_i)\Gamma(\beta_i)}{\Gamma(\alpha_i+\beta_i)}$.

3.2 KL Divergence

Because each component of the random vector is independent, we first find the KL-divergence between two single-variate beta-distribution. Allowing P to be parametrized by a, b and Q by c, d, we find that

$$\begin{aligned} KL(P||Q) &= \int_{[0,1]} [\log\Gamma(a+b) - \log\Gamma(c+d) - \log\Gamma(a) - \log\Gamma(b) + \log\Gamma(c) + \log\Gamma(d) \\ &\quad + (a-c)\log x + (b-d)\log(1-x)]\, dP(x) \\ &= \log\Gamma(a+b) - \log\Gamma(c+d) - \log\Gamma(a) - \log\Gamma(b) + \log\Gamma(c) + \log\Gamma(d) \\ &\quad + (a-c)(\psi^0(a) - \psi^0(a+b)) + (b-d)(\psi^0(b) - \psi^0(a+b)) \end{aligned} \tag{3.1}$$

Now consider two multivariate beta distributions, P and Q, parametrized by α^1, β^1 and α^2, β^2, respectively. The KL-divergence between P and Q is given by

$$\begin{aligned} KL(P||Q) = \sum_{i=1}^{d} &\log\Gamma(\alpha_i^1 + \beta_i^1) - \log\Gamma(\alpha_i^2 + \beta_i^2) - \log\Gamma(\alpha_i^1) - \log\Gamma(\beta_i^1) + \log\Gamma(\alpha_i^2) + \log\Gamma(\beta_i^2) \\ &+ (\alpha_i^1 - \alpha_i^2)(\psi^0(\alpha_i^1) - \psi^0(\alpha_i^1 + \beta_i^1)) + (\beta_i^1 - \beta_i^2)(\psi^0(\beta_i^1) - \psi^0(\alpha_i^1 + \beta_i^1)) \end{aligned} \tag{3.2}$$

3.3 Fisher Information

For notational compactness we typically denote the entry-wise application of a single variate function by overloading its notation – i.e. for $x \in \mathbb{R}^d$ we may write $\log(x) := (\log(x_i))_i$. First we find the differential of l as a function of α and β as

$$\begin{aligned} \mathrm{d}l(\alpha, \beta; x) &= \sum_i \log(x_i)\mathrm{d}\alpha_i + \log(1-x_i)\mathrm{d}\beta_i + \psi^{(0)}(\alpha_i + \beta_i)(\mathrm{d}\alpha_i + \mathrm{d}\beta_i) - \psi^{(0)}(\alpha_i)\mathrm{d}\alpha_i - \psi^{(0)}(\beta_i)\beta_i \\ &= \Big[\underbrace{\log(x) + \psi^{(0)}(\alpha+\beta) - \psi^{(0)}(\alpha)}_{:= g_1(\alpha,\beta)}; \underbrace{\log(1-x) + \psi^{(0)}(\alpha+\beta) - \psi^{(0)}(\beta)}_{:= g_2(\alpha,\beta)}\Big]^\top [\mathrm{d}\alpha; \mathrm{d}\beta]\,. \end{aligned} \tag{3.3}$$

Next we find the differentials of g_1 and g_2 as

$$\begin{aligned} \mathrm{d}g_1(\alpha, \beta) &= \mathrm{diag}[\psi^{(1)}(\alpha+\beta)](\mathrm{d}\alpha + \mathrm{d}\beta) - \mathrm{diag}[\psi^{(1)}(\alpha)]\mathrm{d}\alpha, \\ \mathrm{d}g_2(\alpha, \beta) &= \mathrm{diag}[\psi^{(1)}(\alpha+\beta)](\mathrm{d}\alpha + \mathrm{d}\beta) - \mathrm{diag}[\psi^{(1)}(\beta)]\mathrm{d}\beta. \end{aligned}$$

Plugging into (3.3) the above expressions, we find $\mathrm{d}^2 l(\alpha, \beta)$ as

$$\begin{aligned} \mathrm{d}^2 l(\alpha,\beta) = \mathrm{d}\alpha^\top &\mathrm{diag}[\psi^{(1)}(\alpha+\beta) - \psi^{(1)}(\alpha)]\mathrm{d}\alpha + 2\mathrm{d}\alpha^\top \mathrm{diag}[\psi^{(1)}(\alpha+\beta)]\mathrm{d}\beta \\ &+ \mathrm{d}\beta^\top \mathrm{diag}[\psi^{(1)}(\alpha+\beta) - \psi^{(1)}(\beta)]\mathrm{d}\beta. \end{aligned} \tag{3.4}$$

Applying Lemma A.2 to (3.4) gives

$$\mathrm{H}l(\alpha,\beta) = \begin{bmatrix} \mathrm{diag}[\psi^{(1)}(\alpha+\beta) - \psi^{(1)}(\alpha)] & \mathrm{diag}[\psi^{(1)}(\alpha+\beta)] \\ \mathrm{diag}[\psi^{(1)}(\alpha+\beta)] & \mathrm{diag}[\psi^{(1)}(\alpha+\beta) - \psi^{(1)}(\beta)] \end{bmatrix}.$$

Thus the Fisher Information with respect to the parameters (α, β) is given by

$$\mathcal{I}(\alpha, \beta) = \begin{bmatrix} \mathrm{diag}[\psi^{(1)}(\alpha) - \psi^{(1)}(\alpha+\beta)] & -\mathrm{diag}[\psi^{(1)}(\alpha+\beta)] \\ -\mathrm{diag}[\psi^{(1)}(\alpha+\beta)] & \mathrm{diag}[\psi^{(1)}(\beta) - \psi^{(1)}(\alpha+\beta)] \end{bmatrix}. \tag{3.5}$$

3.4 Modeling the Natural Parameters

We model $\alpha = f(\theta_\alpha)$ and $\beta = g(\theta_\beta)$, giving the parameter $\theta := (\theta_\alpha, \theta_\beta)$. We can write η (defined above) as a function $\eta(\theta)$ Then the Fisher information is given by

$$\mathcal{I}(\theta) = \mathrm{D}_\theta \eta(\theta)^\top \mathcal{I}(\omega) \mathrm{D}_\theta \eta(\theta),$$

where

$$\mathrm{D}_\theta \eta(\theta) = \begin{bmatrix} \mathrm{D}_{\theta_\alpha} f & 0 \\ 0 & \mathrm{D}_{\theta_\beta} g \end{bmatrix}.$$

Combining these expressions and multiplying out gives

$$\mathcal{I}(\theta) = \begin{bmatrix} \mathrm{D}_{\theta_\alpha} f^\top \mathcal{I}(\eta)_{1,1} \mathrm{D}_{\theta_\alpha} f & \mathrm{D}_{\theta_\alpha} f^\top \mathcal{I}(\eta)_{1,2} \mathrm{D}_{\theta_\beta} g \\ \mathrm{D}_{\theta_\beta} g^\top \mathcal{I}(\eta)_{2,1} \mathrm{D}_{\theta_\alpha} f & \mathrm{D}_{\theta_\beta} g^\top \mathcal{I}(\eta)_{2,2} \mathrm{D}_{\theta_\beta} g \end{bmatrix}. \tag{3.6}$$

4 Multivariate Gamma Distribution

If a random variable $X \in (0, \infty)$ is distributed according to the beta distribution with parameters (α, β), then it has the density

$$f(x; \alpha, \beta) = \frac{\beta^\alpha x^{\alpha-1} \exp(-\beta x)}{\Gamma(\alpha)},$$

where $\alpha, \beta > 0$.

We consider the random vector $x \in \mathbb{R}^d$ distributed according to the product d independent beta distributions, each of which is parametrized by some α_i, β_i. In particular,

$$f(x; \alpha, \beta) = \prod_{i=1}^{d} \frac{\beta^\alpha x_i^{\alpha_i - 1} \exp(-\beta_i x_i)}{\Gamma(\alpha_i)}.$$

The log-likelihood function is given by

$$l(\alpha, \beta; x) = \sum_{i=1}^{d} \alpha_i \log(\beta_i) + (\alpha_i - 1)\log(x_i) - \beta_i x_i - \log \Gamma(\alpha_i).$$

4.1 Natural Parametrization

In this section we show that $\eta = [\alpha - 1; -\beta]$ is the natural parameter of the multi-variate gamma distribution. Indeed,

$$\begin{aligned} f(x; \alpha, \beta) &= h(x) \exp\left[\sum_{i=1}^{d} \alpha_i \log(\beta_i) + (\alpha_i - 1)\log(x_i) - \beta_i x_i - \log \Gamma(\alpha_i)\right] \\ &= h(x) \exp\left[T(x)^\top \eta - A(\eta)\right] \end{aligned}$$

where $h(x) = 1$, $T(x) = [\log(x); x]$, and $A(\eta) = \sum_{i=1}^{d} \alpha_i \log \beta_i - \log \Gamma(\alpha_i)$.

4.2 KL Divergence

Because each component of the random vector is independent, we first find the KL-divergence between two single-variate gamma-distribution. Allowing P to be parametrized by a, b and Q by c, d, we find that

$$\begin{aligned}
KL(P||Q) &= \int_{[0,\infty)} [a \log b - c \log d + (a - c) \log x + (d - b)x - \log \Gamma(a) + \log \Gamma(c)]\, dP(x) \\
&= a \log b - c \log d + (a - c)\mathbb{E}_P[\log x] + (d - b)\mathbb{E}_P[x] - \log \Gamma(a) + \log \Gamma(c) \\
&= a \log b - c \log d + (a - c)(\psi^{(0)}(a) - \log b) + (d - b)(a/b) - \log \Gamma(a) + \log \Gamma(c) \\
&= c \log b - c \log d + (a - c)\psi^{(0)}(a) + (d - b)(a/b) - \log \Gamma(a) + \log \Gamma(c), \qquad (4.1)
\end{aligned}$$

where we used that $\mathbb{E}_P[\log x] = \psi^{(0)}(a) - \log b$ and $\mathbb{E}_P[x] = a/b$.

Now consider two multivariate gamma distributions, P and Q, parametrized by α^1, β^1 and α^2, β^2, respectively. The KL-divergence between P and Q is given by

$$\begin{aligned}
KL(P||Q) &= \sum_{i=1}^{d} \alpha_i^2 \log \beta_i^1 - \alpha_i^2 \log \beta_i^2 + (\alpha_i^1 - \alpha_i^2)\psi^{(0)}(\alpha_i^1) + (\beta_i^2 - \beta_i^1)(\alpha_i^1/\beta_i^1) - \log \Gamma(\alpha_i^1) + \log \Gamma(\alpha_i^2) \\
&= \alpha^{2,\top}(\log \beta^1 - \log \beta^2) + (\alpha^1 - \alpha^2)^\top \psi^{(0)}(\alpha^1) + (\beta^2 - \beta^1)^\top(\alpha^1/\beta^1) + (\log \Gamma(\alpha^2) - \log \Gamma(\alpha^1))^\top \mathbf{1}
\end{aligned} \qquad (4.2)$$

4.3 Fisher Information

Recall that $\mathbb{E}_\eta[T(x)] = [\psi^{(0)}(\eta_1 + 1) - \log(-\eta); \alpha \odot \beta^{-1}] := E(\eta)$. To find the Fisher Information, we can use (1.3) by first observing that

$$\begin{aligned}
\mathrm{d}E_1(\eta) &= \mathrm{d}[\psi^{(0)}(\eta_1 + 1)] - \mathrm{d}[\log(-\eta)] \\
&= \mathrm{diag}(\psi^{(0)}(\alpha))\mathrm{d}\eta_1 + \mathrm{diag}(\beta^{-1})\mathrm{d}\eta_2 \qquad (4.3)
\end{aligned}$$

and that

$$\begin{aligned}
\mathrm{d}E_2(\eta) &= \mathrm{d}[\eta_1 + 1] \odot \beta^{-1} - (\eta_1 + 1)\mathrm{d}[-\eta^{-1}] \\
&= \mathrm{diag}(\beta^{-1})\mathrm{d}\eta_1 + \mathrm{diag}(\alpha\beta^{-2})\mathrm{d}\eta_2. \qquad (4.4)
\end{aligned}$$

Combining (4.3) and (4.4) gives

$$\mathcal{I}(\eta) = \begin{bmatrix} \mathrm{diag}(\psi^{(1)}(\alpha)) & \mathrm{diag}(\beta^{-1}) \\ \mathrm{diag}(\beta^{-1}) & \mathrm{diag}(\alpha\beta^{-2}) \end{bmatrix}, \qquad (4.5)$$

or equivalently

$$\mathcal{I}(\omega) = \begin{bmatrix} \mathrm{diag}(\psi^{(1)}(\alpha)) & -\mathrm{diag}(\beta^{-1}) \\ -\mathrm{diag}(\beta^{-1}) & \mathrm{diag}(\alpha\beta^{-2}) \end{bmatrix}, \qquad (4.6)$$

in terms of the parameterization $\omega := [\alpha; \beta]$.

4.4 Modeling the Parameters

We model $\alpha = f(\theta_\alpha)$ and $\beta = g(\theta_\beta)$, giving the parameter $\theta := (\theta_\alpha, \theta_\beta)$. We can write ω (defined above) as a function $\omega(\theta)$. Then the Fisher information is given by

$$\mathcal{I}(\theta) = \mathrm{D}_\theta\omega(\theta)^\top \mathcal{I}(\omega)\mathrm{D}_\theta\omega(\theta),$$

where

$$\mathrm{D}_\theta\omega(\theta) = \begin{bmatrix} \mathrm{D}_{\theta_\alpha} f & 0 \\ 0 & \mathrm{D}_{\theta_\beta} g \end{bmatrix}.$$

Combining these expressions and multiplying out gives

$$\mathcal{I}(\theta) = \begin{bmatrix} \mathrm{D}_{\theta_\alpha} f^\top \mathcal{I}(\omega)_{1,1} \mathrm{D}_{\theta_\alpha} f & \mathrm{D}_{\theta_\alpha} f^\top \mathcal{I}(\omega)_{1,2} \mathrm{D}_{\theta_\beta} g \\ \mathrm{D}_{\theta_\beta} g^\top \mathcal{I}(\omega)_{2,1} \mathrm{D}_{\theta_\alpha} f & \mathrm{D}_{\theta_\beta} g^\top \mathcal{I}(\omega)_{2,2} \mathrm{D}_{\theta_\beta} g \end{bmatrix}. \tag{4.7}$$

5 Natural Policy Gradient Algorithm

To obtain the natural gradient, we need to find a vector that satisfies

$$\mathcal{I}(\theta)g = q_\pi(s, a)\nabla \log f_\pi(a|s).$$

We can do this using the conjugate gradient method (Schulman et al., 2015).

5.1 Details for MV Gaussian with Diagonal Covariance

From Section 2.3 we have a closed form expression for the Fisher Information in (2.18) when the mean is modeled by a function approximator and σ is modeled directly. The only quantity that needs to be found algorithmically is $\mathrm{D}_\theta f$.

Let p denote the dimension of θ. The key operation any implementation needs to provide is the Fisher-vector product $\mathcal{I}(\theta)v$ for arbitrary $v \in \mathbb{R}^{p+d}$. Denoting this operation by F, we have from (2.18) that

$$\begin{aligned} F(v) &= \left[\left(\mathrm{D}_\theta f^\top \mathcal{I}(\omega)_{1,1} \mathrm{D}_\theta f v_1 \right)^\top ; \left(\mathcal{I}(\omega)_{2,2} v_1 \right)^\top \right]^\top \\ &= \left[\left((\mathrm{D}_\theta f^\top)(\mathrm{diag}(\sigma^{-1})\mathrm{D}_\theta f v_1 \right)^\top ; \left(\frac{1}{2}\mathrm{diag}(\sigma^{-2}) v_1 \right)^\top \right]^\top . \end{aligned} \tag{5.1}$$

We can pre-compute $\mathrm{D}_\theta f^\top$ and $\mathrm{diag}(\sigma^{-1})\mathrm{D}_\theta f$, which require $\mathcal{O}(pd)$ memory to store. Because p is usually many orders of magnitude larger than d, this is only marginally more expensive than storing the network itself. To perform all the matrix multiplications requires $\mathcal{O}(pd)$ arithmetic operations.

5.2 Details for Multivariate Beta Distribution

As before, we need to find the fisher-vector product $F(v)$ which from Section 3 we can find as

$$F(v) = \left[\left(\mathrm{D}_\theta f^\top \mathcal{I}(\eta)_{1,1} \mathrm{D}_\theta f v_1 + \mathrm{D}_\theta f^\top \mathcal{I}(\eta)_{1,2} \mathrm{D}_\theta g v_2 \right)^\top ; \left(\mathrm{D}_\theta g^\top \mathcal{I}(\eta)_{2,1} \mathrm{D}_\theta f v_1 + \mathrm{D}_\theta g^\top \mathcal{I}(\eta)_{2,2} \mathrm{D}_\theta g v_2 \right)^\top \right]^\top .$$

Denote by p the dimension of θ and d the dimension of α, β. We can pre-compute the quantities $\mathrm{D}_\theta f^\top \mathcal{I}(\eta)_{1,1}$, $\mathrm{D}_\theta f^\top \mathcal{I}(\eta)_{1,2}$, $\mathrm{D}_\theta g^\top \mathcal{I}(\eta)_{2,1}$, and $\mathrm{D}_\theta g^\top \mathcal{I}(\eta)_{2,2}$. Because all the sub-matrices in $\mathcal{I}$ are diagonal, only $\mathcal{O}(pd)$ operations are required. Likewise, the memory required to store the precomputed quantities is $\mathcal{O}(pd)$.

Using the precomputed quantities, each FVP requires eight matrix vector products requiring $\mathcal{O}(pd)$ operations total. Typically $d \ll p$, so the computational burden is far less than that required to compute the Hessian of the log-likelihood directly which clearly requires $\Omega(d^2)$ operations.

6 Some RL Problems

6.1 Production Problem 1

In this problem we consider an optimal production problem (quite similar to a news-vendor model). Specifically, we model the sequential decision making problem faced by a factory manager who at each time step t can produce goods from a set $\mathcal{G}$. The semantics of the decision making problem are: at time t the manager decides (1) how many of each good to produce subject to having the requisite raw materials and (2) how many of each material $m \in \mathcal{M}$ to order for the next time period; then demand $d \in \mathbb{R}_+^{|\mathcal{G}|}$ is realized and the manager sells as many goods as possible, constrained only by d and his inventory level. For simplicity, we assume that goods and raw materials can be bought, produced and sold in (nonnegative) real valued quantities and that there are no budget constraints. This problem is naturally modeled as a Markov Decision Process (MDP), and below we formulate it as a reward maximization problem.

Formal Model

To formulate the problem, we introduce the following parameters

- $\mathcal{G}$, $\mathcal{M}$ – set of goods that can be produced and raw materials that can be used
- $\mathcal{M}(g)$ – set of materials needed for good g. For all g and g', $\mathcal{M}(g) \cap \mathcal{M}(g') = \varnothing$.
- $a[g, m]$ – amount of material m needed to produce one unit of g
- $p[g]$, $p[m]$ – price of a good $g \in \mathcal{G}$ or material $m \in \mathcal{M}$
- $c[g]$, $c[m]$ – storage cost for good $g \in \mathcal{G}$ or material $m \in \mathcal{M}$ per time period

The state s_t at time t:

- $s_t[g], s_t[m]$ – quantity of good g or material m at start of time period t

The manager makes a decision $u_t \in \mathcal{U}(s_t)$, defined by

- *Production Decision:* $u_t[m]$, $u_t[g]$ – raw materials to purchase and goods to produce at time t.
- *Available Actions:* $\mathcal{U}(s) = \{u : u \geq 0, u[g] \leq l(s)[g] \text{ for all } g \in \mathcal{G}\}$
- *Max. Possible Production:* $l(s)[g] = \min_{m \in \mathcal{M}(g)}\{s[m]/a[g,m]\}$

After observing s_t and making decision u_t, demand is realized, giving r_{t+1} and s_{t+1} as follows:

- *Realized demand:* $d_t[g]$
- *Goods Inventory:* $s_{t+1}[g] = \max\{s_t[g] + u_t[g] - d, 0\}$ for each g
- *Materials Inventory:* $s_{t+1}[m] = s_t[m] - a[g,m]u_t[g] + u_t[m]$ for $g \in \mathcal{G}$ and $m \in \mathcal{M}(g)$
- *Reward:*

$$r_{t+1} = \sum_{g \in \mathcal{G}} p[g] \min\{s_t[g] + u_t[g], d\} - \left[\sum_{m \in \mathcal{M}} p[m]u_t[m] + \sum_{g \in \mathcal{G}} c[g]s_{t+1}[g] + \sum_{m \in \mathcal{M}} c[m]s_{t+1}[m] \right]$$

There are multiple choices for objective, but we consider the finite horizon, cumulative discounted reward maximization problem. Formally, the manager's goal is

$$\max_{\pi \in \Pi} J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_{t+1} \right],$$

where Π is the set of all Markov policies.

RL Formulation

Because the demand distribution is not known a priori, an appropriate choice of stochastic policy is one with support $[0, \infty)$. One viable choice then is the gamma distribution. In keeping with the model free nature of many modern RL approaches, we can allow an agent to interact with the environment by sending actions and receiving observations in $\mathcal{S} = \mathcal{U} = [0, \infty)^{|\mathcal{G}|+|\mathcal{M}|}$. In this way, the agent does not need knowledge of the problem parameters or to directly observe demand at each time step to learn an optimal policy.

Instance 1 – Results

In this first instantiation, we choose the problem parameters

- $\mathcal{G} = \{\text{I}, \text{II}, \text{III}, \text{IV}\}$, $\mathcal{M}(g) = \{\text{g.A}, \text{g.B}, \text{g.B}\}$
- $a[g,m] = 1$ for all g and $m \in \mathcal{M}(g)$
- $p[\text{I}], p[\text{II}], p[\text{III}], p[\text{IV}] = 1.75, 2, 2.25, 2.5$

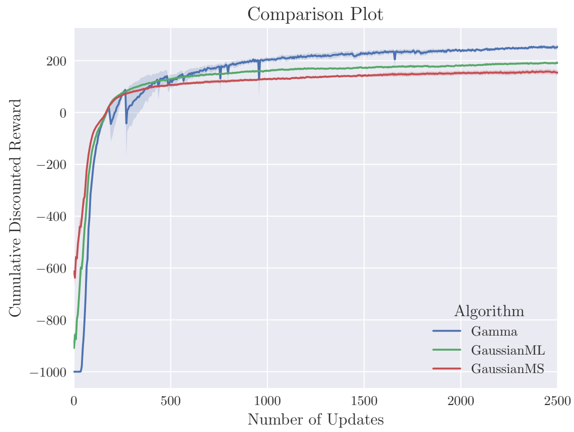

Figure 1: Results on Instance 1 of the Optimal Production problem

- $p[\text{g.A}] = p[\text{g.B}] = p[\text{g.C}] = 0.5$ for all g
- $c[g] = c[m] = 0.1$ for all g and m
- $d[g] \sim \Gamma(k, \mu)$ where $k = 9.0$ is the shape and $\mu = 4.5$ is the mean parameter
- $\gamma = 0.995$, $T = 1000$

We used the same architecture for each parameter – a feed-forward neural network with 2 hidden layers of 32 nodes, with tanh activation functions after each hidden layer. For the Gamma distribution, we model the parameters α and β, passing the final output of each parameter network through a soft-plus activation to ensure it is positive. For the normal distribution, we consider two parameterizations: (1) μ, $\sigma^{1/2}$, the mean and standard deviation, and (2) μ and $\log \sigma^{1/2}$. For the first parametrization, no activation is applied to the output for μ and a soft-plus is applied to the parameter net for $\sigma^{-1/2}$. For the second parameterization, no activation is applied to the output of either parameter net. Figure 1 compares the performance of TRPO using each of the distributions. In the plot, the Gaussian parameterizations are labeled GaussianMS and GaussianML, respectively. Using the Gamma distribution outperforms the Gaussian by a factor of 25%.

References

Chou, P.-W., Maturana, D. and Scherer, S. (2017). Improving Stochastic Policy Gradients in Continuous Control with Deep Reinforcement Learning using the Beta Distribution. In *ICML*.

Eisenach, C., Yang, H., Liu, J. and Liu, H. (2018). Marginal Policy Gradients for Complex Control. arXiv:1806.05134.

Fujita, Y. and Maeda, S.-I. (2018). Clipped Action Policy Gradient. In *ICML*.

Schulman, J., Levine, S., Moritz, P., Jordan, M. and Abbeel, P. (2015). Trust Region Policy Optimization. In *ICML*.

Schulman, J., Moritz, P., Levine, S., Jordan, M. and Abbeel, P. (2016). High-Dimensional Continuous Control Using Generalized Advantage Estimation. In *ICLR*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature* **529** 484–489.

A Preliminary Results

A.1 The vec operator and the Kronecker Product

Several useful properties of the Kronecker product and vec operator are given below. Notationally, lower case letters indicate vectors and uppercase letters, matrices.

$$\begin{aligned}\operatorname{vec}(ABC) &= \left(C^\top \otimes A\right) \operatorname{vec} B,\\ A\left(b^\top \otimes I_{d,d}\right) &= b^\top \otimes A,\\ \left(b \otimes I_{d,d}\right) A &= b \otimes A\end{aligned}$$

A.2 Results in Differential Calculus

A.2.1 Partitioning the Hessian

In this section we demonstrate some useful lemmas for finding the Hessian of a function of a partitioned vector.

Lemma A.1. Let $\phi : \mathbb{R}^{p+q} \to \mathbb{R}$ be a twice differentiable scalar function. Let $x \in \mathbb{R}^{p+q}$ be partitioned as $x = [x_1^\top; x_2^\top]^\top$ where $x_1 \in \mathbb{R}^p$ and $x_2 \in \mathbb{R}^q$. If $\mathrm{d}^2\phi(x) = \mathrm{d}x_1^\top A \mathrm{d}x_2$ for some $A \in \mathbb{R}^{p\times q}$, then

$$\mathrm{H}_x\phi(x) = \begin{bmatrix} 0_{p,p} & \frac{1}{2}A \\ \frac{1}{2}A^\top & 0_{q,q}\end{bmatrix}.$$

Proof. First, observe that we can write

$$x_1 = [I_p; 0_{p,q}]\, x, \quad \text{and} \quad x_2 = [0_{q,p}; I_{q,q}]\, x.$$

Then it follows that

$$\begin{aligned}\mathrm{d}x_1^\top A \mathrm{d}x_2 &= \mathrm{d}x^\top\, [I_p; 0_{p,q}]^\top\, A\, [0_{q,p}; I_{q,q}]\, \mathrm{d}x\\ &= \mathrm{d}x^\top \left[A^\top; 0_{q,q}\right]^\top [0_{q,p}; I_{q,q}]\, \mathrm{d}x\\ &= \mathrm{d}x^\top \begin{bmatrix} 0_{p,p} & A \\ 0_{q,p} & 0_{q,q}\end{bmatrix} \mathrm{d}x.\end{aligned}$$

The result now follows immediately from the identification theorem for the second differential. □

Lemma A.2. Let $\phi : \mathbb{R}^{p+q} \to \mathbb{R}$ be a twice differentiable scalar function. Let $x \in \mathbb{R}^{p+q}$ be partitioned as $x = [x_1^\top; x_2^\top]^\top$ where $x_1 \in \mathbb{R}^p$ and $x^2 \in \mathbb{R}^q$. If

$$\mathrm{d}^2\phi(x) = \mathrm{d}x_1^\top A \mathrm{d}x_1 + \mathrm{d}x_1^\top B \mathrm{d}x_2 + \mathrm{d}x_2^\top C \mathrm{d}x_2$$

for some $A \in \mathbb{R}^{p\times p}$, $B \in \mathbb{R}^{p\times q}$, and $C \in \mathbb{R}^{q\times q}$ then

$$\mathrm{H}_x\phi(x) = \frac{1}{2}\begin{bmatrix} (A+A^\top) & B \\ B^\top & (C+C^\top)\end{bmatrix}.$$

Proof. Following an analogous argument to the one made in the proof of Lemma A.1, we obtain

$$\mathrm{d}x_1^\top A \mathrm{d}x_1 + \mathrm{d}x_1^\top B \mathrm{d}x_2 + \mathrm{d}x_2^\top C \mathrm{d}x_2 = \mathrm{d}x^\top \begin{bmatrix} A & B \\ 0_{q,p} & C \end{bmatrix} \mathrm{d}x.$$

After symmetrizing, the result follows from the identification theorem for the second differential. $\square$