# Projections for Approximate Policy Iteration Algorithms: Supplementary

**Anonymous Authors**[1]

## Appendix A. Projections for discrete distributions

We have introduced so far a set of projections to help solve optimization problems involving Gaussian distributions. The projections allow one to tackle the maximization of an objective function under entropy and (I-projection) KL constraint. While finding an appropriate projection for a constraint is not trivial, the advantage of our approach to constrained optimization is that the projection is independent of the objective and can be used to optimize any objective function. We discuss in this section how projections can be used to add entropy and KL constraints to discrete action RL algorithms.

For discrete action spaces, a usual choice is for $\pi$ to be a soft-max distribution $\pi(a_i|s) \propto \exp(f^i_\omega(s))$ where $f^i_\omega$ is the i-th output of parameterized function $f_\omega$. From here on we term $f_\omega$ the 'logits' of $\pi$, and let $\mathcal{H}(f_\omega(s))$ be the entropy of the associated soft-max distribution. For a given $s$, let $r_i$ be the probability of action $i$ according to $f_\omega$, i.e. $r_i \propto \exp(f^i_\omega(s))$. To ensure satisfaction of the entropy constraint, we derive a projection $g_\beta$ such that $\mathcal{H}(g_\beta \circ f_\omega(s)) \geq \beta$ for all $s$. The resulting policy $\pi$ of logits $g_\beta \circ f_\omega$ is given by

$$\pi(a_i|s) = \begin{cases} r_i, & \text{if } \mathcal{H}(f_\omega) \geq \beta \\ \alpha r_i + (1-\alpha)\frac{1}{|\mathcal{A}|}, & \text{otherwise} \end{cases}$$

where $\alpha = \frac{\log(|\mathcal{A}|)-\beta}{\log(|\mathcal{A}|)-\mathcal{H}(f_\omega)}$. This policy will always comply with the constraint $\mathcal{H}(\pi(.|s)) \geq \beta$ for all $s$. It is true by definition for $\mathcal{H}(f_\omega) \geq \beta$ and can easily be verified when $\mathcal{H}(f_\omega) < \beta$ since

$$\mathcal{H}\left(\alpha r + (1-\alpha)\frac{1}{|\mathcal{A}|}\right) \geq \alpha\mathcal{H}(f_\omega) + (1-\alpha)\log(|\mathcal{A}|), \tag{1}$$

$$= \beta.$$

The inequality follows from the fact that the entropy of a mixture is greater than the mixture of entropies (Cover & Thomas, 2006). The mixture being between the probability distribution defined by $r$ which has entropy $\mathcal{H}(f_\omega)$ and the uniform distribution which has entropy $\log(|\mathcal{A}|)$. The equality to $\beta$ in the above set of equations follows from the definition of $\alpha$. This projection is for the constraint $\mathcal{H}(g_\beta \circ f_\omega(s)) \geq \beta$ for all $s$. If one desires to have $\mathbb{E}_s[\mathcal{H}(g_\beta \circ f_\omega(s))] \geq \beta$ instead, applying the expectation to the lower bound of inequality (1) indicates that the appropriate interpolation parameter is obtained by replacing $\mathcal{H}(f_\omega)$ by its expectation over $s$ in the definition of $\alpha$. The projection for the KL follows a similar principle, simply replacing the uniform distribution in the interpolation with $q$, and using the same argument of mixture to obtain a linear equation in the interpolation parameter.

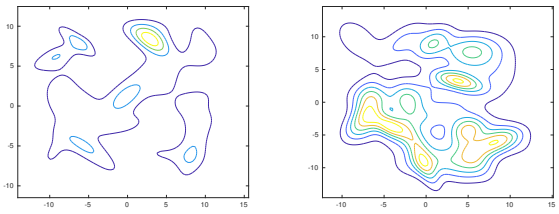

*Figure 1.* Sampled objective functions to be optimized by the DPS algorithms.

## Appendix B. DPS experiment

We evaluate Alg. 1 for solving DPS search distribution update introduced in Sec. 2.1. The algorithm is tasked to optimize randomly generated smooth functions, two sample of which are shown in Fig. 1. We compare the optimization of $L \circ g$ with $g$ as defined in Alg. 1 to two baselines from the literature relying on the method of Lagrange multipliers.

### State-of-the-art baselines

The considered baseline algorithms to our projection method are REPS (Peters et al., 2010) and MORE (Abdolmaleki et al., 2015) that both rely on the method of Lagrange multipliers to obtain a closed form solution to the constrained problem. REPS (Peters et al., 2010; Deisenroth et al., 2013) solves the same DPS update problem bar the entropy constraint which it lacks. The closed form solution of the update

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

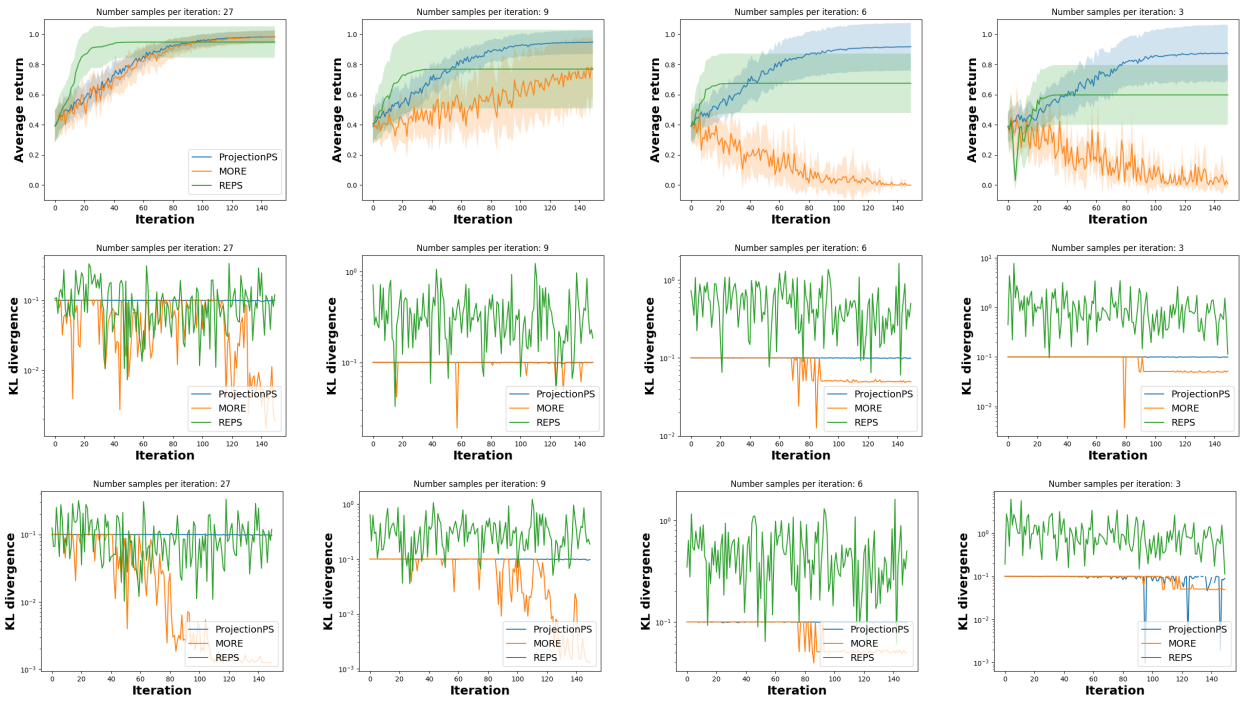


*Figure 2.* Optimization of smooth objective functions with varying number of samples per iteration, with values of 27, 9, 6, and 3 from left to right columns. First row shows the average return at each iteration averaged over 11 runs. Second and third row show the KL divergence between successive policies of two runs.

is given by

$$\pi(\theta) \propto q(\theta) \exp\left(\frac{\bar{R}(\theta)}{\eta^*}\right),$$

where $\bar{R}(\theta) = \mathbb{E}[R(\theta)]$ and $\eta^*$ is a dual variable that is computed using gradient descent. However, $\pi$ is not necessarily Gaussian and an additional weighted maximum likelihood step is necessary to obtain the next search distribution. This final step can cause large violations of the KL constraint.

MORE (Abdolmaleki et al., 2015) solves the same DPS update problem, but uses $\hat{R}$, a quadratic approximation of $R$ learned by linear regression. The resulting policy is

$$\pi(\theta) \propto q(\theta)^{\eta^*/(\eta^*+\omega^*)} \exp\left(\frac{\hat{R}(\theta)}{\eta^*+\omega^*}\right).$$

As $\hat{R}$ is quadratic and $q$ Gaussian the resulting search distribution remains Gaussian and the KL and entropy constraints are never violated.

## Experiment

We compare our approach to the two baselines of the previous section for the optimization of randomly generated and smooth two dimensional objective functions, illustrated in Fig. 1. The results are reported in Fig. 2 on 11 independent runs and varying number of samples per iteration. The 11 randomly generated functions are sampled once and kept fixed for all the algorithms and varying hyper-parameters. For each function, the reported results are mapped to $[0, 1]$ after computing the minimal and maximal values reached for this function across all algorithms and hyper-parameters.

First row of Fig. 2 shows the average return at each iteration for the three direct policy search algorithms. The number of samples per iteration takes values 27, 9, 6 and 3 from left to right column respectively while the dimensionality of the problem is $d = 2$. Our approach, termed 'ProjectionPS' is very robust to reduction in sample count and changes moderately across scenarios. While REPS exhibits signs of premature convergence as the sample count drops, caused by large KL constraint violations as seen in Fig. 2, second and third row. MORE never violates the KL constraint but the quadratic models are of poor quality using only 3 and 6 samples which deteriorates performance. Our algorithm nearly always returns a solution with maximum allowed KL constraint $\epsilon = .1$ apart from a single run with a sample count of 3 as seen in Fig. 2.

## Appendix C. Extended proofs of the propositions

We extend the proofs of the propositions of the main paper in this appendix. Starting with the entropy projection, we recall the definition of $h$,

$$h(\lambda, c) = \left(\frac{d}{2}\log(2\pi e) + \sum_i \lambda_i\right) - \beta. \tag{2}$$

**Proposition 1.** *Optimizing any function $L(\pi)$ w.r.t. mean vector $\mu$ and diagonal matrix $\Sigma$ of a Gaussian $\pi = \mathcal{N}(\mu, \Sigma)$ under entropy equality constraint $\mathcal{H}(\pi) = \beta$ is equivalent to the unconstrained optimization of $L(\pi)$ w.r.t. mean vector $\mu$ and the real valued parameter vector $\lambda$ such that $\Sigma_{i,i} = \exp^2(\lambda_i - \frac{1}{d}h(\lambda, \beta))$ with $h$ as define in Eq. (2).*

*Proof.* We show that any value of $\lambda$ will yield a Gaussian distribution that satisfies the entropy equality constraint and that for any Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ satisfying the entropy constraint there is a parameter vector $\lambda$ representing it. The first implication is verified through straightforward computation, using the definition of $h$ to conclude that the resulting Gaussian has entropy of exactly $\beta$; while for any $\Sigma$ such that $\mathcal{H}(\Sigma) = \beta$, setting $\lambda_i = \frac{1}{2}\log(\Sigma_{i,i})$ will yield back $\Sigma$ since $h(\lambda, \beta) = 0$. Hence optimizing $L(\pi)$ w.r.t. $\Sigma$ under constraint $\mathcal{H}(\pi) = \beta$ is equivalent to the unconstrained optimization of $L(\pi)$ w.r.t. $\lambda$. □

**Proposition 2.** *Optimizing any function $L(\pi)$ w.r.t. mean vector $\mu$ and diagonal matrix $\Sigma$ of a Gaussian $\pi = \mathcal{N}(\mu, \Sigma)$, under entropy inequality constraint $\mathcal{H}(\pi) \geq \beta$ is equivalent to the unconstrained optimization of $L(\pi)$ w.r.t. mean vector $\mu$ and the real valued parameter vector $\lambda$ such that $\Sigma_{i,i} = \exp(2\max(\lambda_i, \lambda_i - \frac{1}{d}h(\lambda, \beta)))$ with $h$ defined in Eq. (2).*

*Proof.* If $\Sigma_{i,i} = \exp(2\max(\lambda_i, \lambda_i - \frac{1}{d}h(\lambda, \beta)))$ and $\Sigma'_{i,i} = \exp(2\lambda_i - \frac{1}{d}h(\lambda, \beta))$ then $\mathcal{H}(\Sigma) \geq \mathcal{H}(\Sigma')$ and we have shown that $\mathcal{H}(\Sigma') = \beta$. Now let a diagonal Gaussian distribution $\pi = \mathcal{N}(\mu, \Sigma)$ such that $\mathcal{H}(\Sigma) \geq \beta$ and let $\lambda$ be the parameter vector such that $\lambda_i = \frac{1}{2}\log(\Sigma_{i,i})$, then $h(\lambda, \beta) \geq 0$ implying that $\max(\lambda_i, \lambda_i - \frac{1}{d}h(\lambda, \beta)) = \lambda_i$. Hence the parameter vector $\lambda$ will yield $\Sigma$. As a result, optimizing $L(\pi)$ w.r.t. $\Sigma$ under constraint $\mathcal{H}(\pi) \geq \beta$ is equivalent to the unconstrained optimization of $L(\pi)$ w.r.t. $\lambda$. □

**Proposition 3.** *Optimizing any function $L(\pi)$ w.r.t. mean vector $\mu$ and covariance $\Sigma$ of a Gaussian $\pi = \mathcal{N}(\mu, \Sigma)$, under entropy inequality constraint $\mathcal{H}(\pi) \geq \beta$ and KL constraint $\mathrm{KL}(\pi \parallel q) \leq \epsilon$ for Gaussian $q$ such that $\mathcal{H}(q) \geq \beta$ is equivalent to the unconstrained optimization of $L(\pi)$ w.r.t. the parameterization given by Alg. 1.*

**Algorithm 1** DPS Gaussian policy projection

**Input:** $\mu$, $\lambda$, $\lambda_{\text{off\_diag}}$, $q = \mathcal{N}(\mu_q, \Sigma_q)$, $\epsilon$ and $\beta$
**Output:** $\pi = \mathcal{N}(\mu, \Sigma)$ complying with KL and entropy constraints
$\Sigma$ = Entropy_projection($\lambda$, $\lambda_{\text{off\_diag}}$, $\beta$)
**if** $\mathrm{KL}(\mathcal{N}(\mu, \Sigma) \parallel q) > \epsilon$ **then**
  $\eta_g = \frac{\epsilon}{m_q(\mu) + r_q(\Sigma) + e_q(\Sigma)}$
  $\Sigma = \eta_g \Sigma + (1 - \eta_g)\Sigma_q$
**end if**
**if** $\mathrm{KL}(\mathcal{N}(\mu, \Sigma) \parallel q) > \epsilon$ **then**
  $\eta_m = \sqrt{\frac{\epsilon - r_q(\Sigma) - e_q(\Sigma)}{m_q(\mu)}}$
  $\mu = \eta_m \mu + (1 - \eta_m)\mu_q$
**end if**

*Proof.* The assumption that $\mathcal{H}(q) \geq \beta$ ensures that the optimization problem admits a valid solution that satisfies both KL and entropy constraint. Let us first show that Alg. 1 returns Gaussian distributions satisfying both constraints irrespective of the input values. Alg. 1 starts by using the entropy projection which from Prop. 2 will result in $\mathcal{H}(\Sigma) \geq \beta$. The remainder of the algorithm simply interpolates the current covariance and mean with that of $q$ to ensure that $\mathrm{KL}(\pi \parallel q) \leq \epsilon$. Letting $\Sigma_\eta = \eta\Sigma + (1 - \eta)\Sigma_q$, for $\eta \in [0, 1]$, the value of $\eta_g$ and $\eta_m$ are derived by trying to upper bound $r_q + e_q$ and $m_q$ respectively. For $e_q(\Sigma_\eta)$

$$\begin{aligned} |\Sigma_\eta|^{\frac{1}{d}} &\geq |\eta\Sigma|^{\frac{1}{d}} + |(1-\eta)\Sigma_q|^{\frac{1}{d}}, \\ &\qquad \text{(Minkowski determinant inequality)} \\ \log|\Sigma_\eta| &\geq \eta\log|\Sigma| + (1-\eta)\log|\Sigma_q|, \\ &\qquad \text{(concavity of log)} \\ e_q(\Sigma_\eta) &\leq \eta e_q(\Sigma). \end{aligned}$$

Exploiting linearity of the trace operator, one can straighforwardly show the same property for $r_q(\Sigma_\eta)$. As a result we have that $r_q(\Sigma_\eta) + e_q(\Sigma_\eta) \leq \eta(r_q(\Sigma) + e_q(\Sigma))$. Note that the entropy constraint is satisfied by $\Sigma_\eta$ for any $\eta \in [0, 1]$ since the second inequality derived from the concavity of the log shows that the entropy of $\Sigma_\eta$ cannot be lower than the entropy of the covariances it interpolates. Similarly for the mean, letting $\mu_\eta = \eta\mu + (1 - \eta)\mu_q$ for $\eta \in [0, 1]$, we have $m_q(\mu_\eta) = \eta^2 m_q(\mu)$. As a result, using the property that the KL is non-negative—which implies non-negativity of $m_q$ and $r_q + e_q$—one can verify that $\eta_g$ and $\eta_m$ are both in $[0, 1]$ and by direct computation using the value of $\eta_g$ and $\eta_m$ in the above inequality and equality, that Alg. 1 returns a distribution satisfying both KL and entropy constraint. Conversely, if $\mathcal{N}(\mu, \Sigma)$ satisfies the KL constraint then it will be unaltered by the KL projection part of Alg. 1 while we know from Prop. 2 that there is a set of parameters to represent any $\Sigma$ satisfying the entropy constraint. □

**Proposition 4.** *Optimizing any function $L(\pi)$ w.r.t. parameters $A'$ and $\Sigma$ of linear in feature Gaussian pol-*

**Algorithm 2** API linear-Gaussian policy projection

**Input:** $A'$, $\lambda$, $\lambda_{\text{off\_diag}}$, $q(.|s) = (A_q^T\psi_q(s), \Sigma_q)$, $A^T$, $\psi$, $\epsilon$ and $\beta$
**Output:** $\pi(.|s) = \mathcal{N}(A'^T\psi(s), \Sigma)$ complying with KL and entropy constraints
$\Sigma$ = Entropy_projection$(\lambda, \lambda_{\text{off_diag}}, \beta)$
**if** $\mathbb{E}_s\text{KL}(\mathcal{N}(A'^T\psi(s), \Sigma) \parallel q(.|s)) > \epsilon$ **then**
  $\eta_g = \frac{\epsilon - m_q(A)}{m_q(A') + r_q(\Sigma) + e_q(\Sigma)}$
  $\Sigma = \eta_g\Sigma + (1 - \eta_g)\Sigma_q$
**end if**
**if** $\mathbb{E}_s\text{KL}(\mathcal{N}(A'^T\psi(s), \Sigma) \parallel q(.|s)) > \epsilon$ **then**
  $a = .5\mathbb{E}_s||A'^T\psi(s) - A^T\psi(s)||^2_{\Sigma_q^{-1}}$
  $b = .5\mathbb{E}_s[(A'^T\psi(s) - A^T\psi(s))^T \Sigma_q^{-1}(A^T\psi(s) - A_q^T\psi_q(s))]$
  $c = m_q(A) + r_q(\Sigma) + e_q(\Sigma) - \epsilon$
  $\eta_m = \frac{-b + \sqrt{b^2 - ac}}{a}$
  $A' = \eta_m A' + (1 - \eta_m)A$
**end if**

*icy* $\pi(.|s) = \mathcal{N}(A'^T\psi(s), \Sigma)$*, under entropy constraint* $\mathbb{E}_{s\sim q}[\mathcal{H}(\pi(.|s))] \geq \beta$ *and KL constraint* $\mathbb{E}_{s\sim q}[\text{KL}(\pi(.|s) \parallel q(.|s))] \leq \epsilon$ *to linear in feature Gaussian policy* $q(.|s) = \mathcal{N}(A_q^T\psi_q(s), \Sigma_q)$ *such that i)* $\mathcal{H}(q) \geq \beta$ *and ii) there exist* $A$ *such that* $m_q(A) \leq \epsilon$*, is equivalent to the unconstrained optimization of* $L(\pi)$ *w.r.t. the parameterization given by Alg. 2.*

*Proof.* The additional assumption (ii) on $q$ compared to Prop. 3 ensures that the feature change from $\psi_q$ to $\psi$ does not preclude the existence of a solution to the optimization problem. In Alg. 2 the mean projection requires an $\eta_m$ such that $m_q\left(\eta_m A' + (1 - \eta_m)A\right) + r_q(\Sigma) + e_q(\Sigma) = \epsilon$ in case of KL violation, i.e. to solve $f(\eta) = a\eta^2 + 2b\eta + c = 0$ with coefficients given in Alg. 2. The solution is given by $\eta_m \in [0, 1]$ as defined in Alg. 2. Indeed, $f(0) \leq 0$ from the definition of $\eta_g$ and $f(1) > 0$ because the KL is violated and since $f$ is continuous and convex ($a \geq 0$) then the quadratic function $f$ accepts a root in $[0, 1]$ and is given by the greater root as in Alg. 2. The rest of the proof follows as for Prop. 3. □

## References


Abdolmaleki, A., Lioutikov, R., Peters, J., Lau, N., Pualo Reis, L., and Neumann, G. Model-based relative entropy stochastic search. In *Advances in Neural Information Processing Systems (NIPS)*. 2015.

Cover, T. M. and Thomas, J. A. *Elements of Information Theory*. Wiley-Interscience, 2006. ISBN 0471241954.

Deisenroth, M. P., Neumann, G., and Peters, J. A Survey on Policy Search for Robotics. *Foundations and Trends in Robotics*, pp. 388–403, 2013.

Peters, J., Mülling, K., and Altün, Y. Relative entropy policy search. In *National Conference on Artificial Intelligence (AAAI)*, 2010.