APPENDICES

A. LSVI with Boltzmann exploration/ϵ-greedy exploration

The LSVI algorithm iterates backwards over time periods in the planning horizon, in each iteration fitting a value function to the sum of immediate rewards and value estimates of the next period. Each value function is fitted via least-squares: note that vectors θ_{lh} satisfy

$$\theta_{lh} \in \operatorname*{arg\,min}_{\zeta \in \mathbb{R}^K} \left(\|A\zeta - b\|^2 + \lambda \|\zeta\|^2 \right). \tag{3}$$

Notice that in Algorithm 3, when $l = 0$, matrix A and vector b are empty. In this case, we simply set $\theta_{l0} = \theta_{l1} = \cdots = \theta_{l,H-1} = 0$.

Algorithm 3 Least-Squares Value Iteration

```
Input: Data $\Phi(s_{i0},a_{i0}),r_{i0},..,\Phi(s_{iH-1},a_{iH-1}),r_{iH} : i < L$
       Parameter $\lambda > 0$
Output: $\theta_{l0},\ldots,\theta_{l,H-1}$
$\theta_{lH} \leftarrow 0$, $\Phi_H \leftarrow \mathbf{0}$
for $h = H-1,\ldots,1,0$ do
    Generate regression problem $A \in \mathbb{R}^{l \times K}$, $b \in \mathbb{R}^l$:
        $A \leftarrow \begin{bmatrix} \Phi_h(s_{0h},a_{0h}) \\ \vdots \\ \Phi_h(s_{l-1,h},a_{l-1,h}) \end{bmatrix}$
        $b_i \leftarrow \begin{cases} r_{ih} + \max_\alpha \left( \Phi_{h+1} \tilde{\theta}_{l,h+1} \right)(s_{i,h+1},\alpha) & \text{if } h < H-1 \\ r_{ih} + r_{i,h+1} & \text{if } h = H-1 \end{cases}$
    Linear regression for value function
        $\theta_{lh} \leftarrow (A^\top A + \lambda I)^{-1} A^\top b$
end for
```

RL algorithms produced by synthesizing Boltzmann exploration or ϵ-greedy exploration with LSVI are presented as Algorithms 4 and 5. In these algorithms the "temperature" parameters η in Boltzmann exploration and ϵ in ϵ-greedy exploration control the degree to which random perturbations distort greedy actions.

Algorithm 4 LSVI with Boltzmann exploration

```
Input: Features $\Phi_0,..,\Phi_{H-1}$; $\eta > 0$, $\lambda > 0$
for $l = 0,1,\cdots$ do
    Compute $\theta_{l0},\ldots,\theta_{l,H-1}$ based on Algorithm 3
    Observe $x_{l0}$
    for $h = 0,1,\ldots,H-1$ do
        Sample $a_{lh} \sim \mathbb{E}[(\Phi_h \theta_{lh})(x_{lh},a)/\eta]$
        Observe $r_{lh}$ and $x_{l,h+1}$
    end for
end for
```

Algorithm 5 LSVI with ϵ-greedy exploration

```
Input: Features $\Phi_0,..,\Phi_{H-1}$; $\epsilon > 0$, $\lambda > 0$
for $l = 0,1,\ldots$ do
    Compute $\theta_{l0},\ldots,\theta_{l,H-1}$ using Algorithm 3
    Observe $x_{l0}$
    for $h = 0,1,\cdots,H-1$ do
        Sample $\xi \sim$ Bernoulli$(\epsilon)$
        if $\xi = 1$ then
            Sample $a_{lh} \sim$ unif$(\mathcal{A})$
        else
            Sample $a_{lh} \in \operatorname{argmax}_{\alpha \in \mathcal{A}} (\Phi_h \theta_{lh})(x_{lh},\alpha)$
        end if
        Observe $r_{lh}$ and $x_{l,h+1}$
    end for
end for
```

B. Efficient exploration with generalization

Our computational results suggest that, when coupled with generalization, RLSVI enjoys levels of efficiency far beyond what can be achieved by Boltzmann or ϵ-greedy exploration. We leave as an open problem establishing efficiency guarantees in such contexts. To stimulate thinking on this topic, we put forth a conjecture.

Conjecture 1. *For all $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, R, \pi)$, $\Phi_0, \ldots, \Phi_{H-1}$, σ, and λ, if reward distributions R have support $[-\sigma, \sigma]$, there is a unique $(\theta_0, \ldots, \theta_{H-1}) \in \mathbb{R}^{K \times H}$ satisfying $Q^*_h = \Phi_h \theta_h$ for $h = 0, \ldots, H-1$, and $\sum_{h=0}^{H-1} \|\theta_h\|^2 \leq \frac{KH}{\lambda}$, then there exists a polynomial* poly *such that*

$$\mathrm{Regret}(T, \mathcal{M}) \leq \sqrt{T}\, \mathrm{poly}\left(K, H, \max_{h,x,a} \|\Phi_h(x,a)\|, \sigma, 1/\lambda \right).$$

As one would hope for from an RL algorithm that generalizes, this bound does not depend on the number of states or actions. Instead, there is a dependence on the number of basis functions. In Appendix C we present empirical results that are consistent with this conjecture.

C. Chain experiments

C.1. Generating a random coherent basis

We present full details for Algorithm 6, which generates the random coherent basis functions $\Phi_h \in \mathbb{R}^{SA \times K}$ for $h = 1, .., H$. In this algorithm we use some standard notation for indexing vector elements. For any $A \in \mathbb{R}^{m \times n}$ we will write $A[i, j]$ for the element in the i^{th} row and j^{th} column. We will use the placeholder $\cdot$ to repesent the entire axis so that, for example, $A[\cdot, 1] \in \mathbb{R}^n$ is the first column of A.

Algorithm 6 Generating a random coherent basis

```
Input: $S, A, H, K \in \mathbb{N}$,  $Q_h^* \in \mathbb{R}^{SA}$ for $h = 1, .., H$
Output: $\Phi_h \in \mathbb{R}^{SA \times K}$ for $h = 1, .., H$
 1: Sample $\Psi \sim N(0, I) \in \mathbb{R}^{HSA \times K}$
 2: Set $\Psi[\cdot, 1] \leftarrow \mathbb{1}$
 3: Stack $Q^* \leftarrow (Q_1^*, .., Q_h^*) \in \mathbb{R}^{HSA}$
 4: Set $\Psi[\cdot, 2] \leftarrow Q^*$
 5: Form projection $P \leftarrow \Psi(\Psi^T \Psi)^{-1} \Psi^T$
 6: Sample $W \sim N(0, I) \in \mathbb{R}^{HSA \times K}$
 7: Set $W[\cdot, 1] \leftarrow \mathbb{1}$
 8: Project $W_P \leftarrow PW \in \mathbb{R}^{HSA \times K}$
 9: Scale $W_P[\cdot, k] \leftarrow \frac{W_P[\cdot, k]}{\|W_P[\cdot, k]\|_2} HSA$ for $k = 1, .., K$
10: Reshape $\Phi \leftarrow \text{reshape}(W_P) \in \mathbb{R}^{H \times SA \times K}$
11: Return $\Phi[h, \cdot, \cdot] \in \mathbb{R}^{SA \times K}$ for $h = 1, .., H$
```

The reason we rescale the value function in step (9) of Algorithm 6 is so that the resulting random basis functions are on a similar scale to Q^*. This is a completely arbitrary choice as any scaling in Φ can be exactly replicated by similar rescalings in λ and σ.

C.2. Robustness to λ, σ

In Figures 10 and 11 we present the cumulative regret for $N = 50, K = 10$ over the first 10000 episodes for several orders of magnitude for σ and λ. For most combinations of parameters the learning remains remarkably stable.

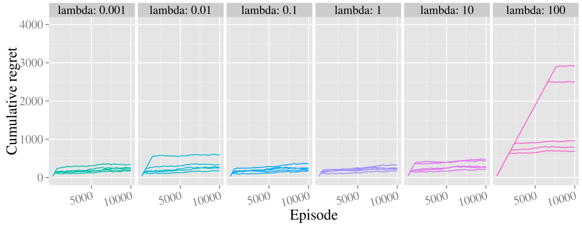

Figure 10. Fixed $\sigma = 0.1$, varying λ.

We find that large values of σ lead to slowers learning, since the Bayesian posterior concentrates only very slowly with new data. However, in stochastic domains we found that choosing a σ which is too small might cause the RLSVI posterior to concentrate too quickly and so fail to sufficiently explore. This is a similar insight to previous analyses of Thompson sampling (Agrawal & Goyal, 2012) and matches the flavour of Theorem 1.

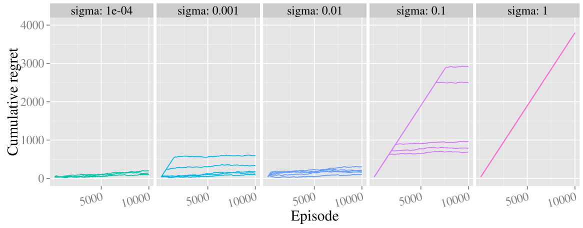

Figure 11. Fixed $\lambda = 100$, varying σ.

C.3. Scaling with number of bases K

In Figure 4 we demonstrated that RLSVI seems to scale gracefully with the number of basis features on a chain of length $N = 50$. In Figure 13 we reproduce these reults for chains of several different lengths. To highlight the overall trend we present a local polynomial regression for each chain length.

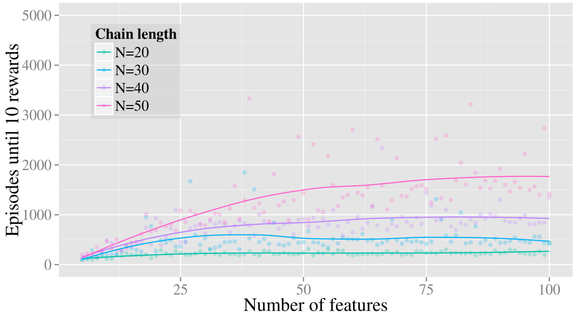

Figure 12. Graceful scaling with number of basis functions.

Roughly speaking, for low numbers of features K the number of episodes required until learning appears to increase linearly with the number of basis features. However, the marginal increase from a new basis features seems to decrease and almost plateau once the number of features reaches the maximum dimension for the problem $K \geq SA$.

C.4. Approximate polynomial learning

Our simulation results empirically demonstrate learning which appears to be polynomial in both N and K. Inspired by the results in Figure 5, we present the learning times for different N and K together with a quadratic regression fit separately for each K.

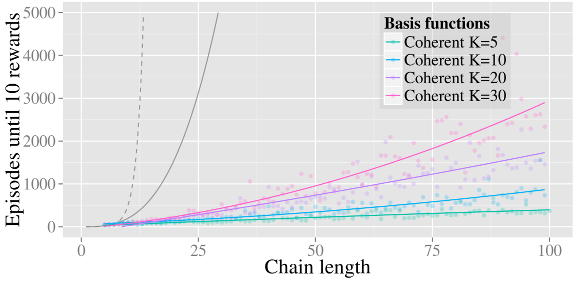

Figure 13. Graceful scaling with number of basis functions.

This is only one small set of experiments, but these results are not inconsistent with Conjecture 1. This quadratic model seems to fit data pretty well.

D. Tetris experiments

D.1. Algorithm specification

In Algorithm 7 we present a natural adaptation to RLSVI without known episode length, but still a regular episodic structure. This is the algorithm we use for our experiments in Tetris. The LSVI algorithms are formed in the same way.

```
Algorithm 7 Stationary RLSVI
Input: Data $\Phi(s_1,a_1),r_1,..,\Phi(s_T,a_T)$
       Previous estimate $\tilde{\theta}_l^- \equiv \tilde{\theta}_{l-1}$
       Parameters $\lambda > 0$, $\sigma > 0$, $\gamma \in [0,1]$
Output: $\tilde{\theta}_l$
  Generate regression problem $A \in \mathbb{R}^{T \times K}$, $b \in \mathbb{R}^T$:
      $A \leftarrow \begin{bmatrix} \Phi_h(s_1,a_1) \\ \vdots \\ \Phi_h(s_T,a_T) \end{bmatrix}$
      $b_i \leftarrow \begin{cases} r_i + \gamma \max_\alpha \left(\Phi \tilde{\theta}_l^-\right)(s_{i+1},\alpha) & \text{if } s_i \text{ not terminal} \\ r_i & \text{if } s_i \text{ is terminal} \end{cases}$
  Bayesian linear regression for the value function
      $\overline{\theta}_l \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}A^\top A + \lambda I\right)^{-1} A^\top b$
      $\Sigma_l \leftarrow \left(\frac{1}{\sigma^2}A^\top A + \lambda I\right)^{-1}$
  Sample $\tilde{\theta}_l \sim N(\overline{\theta}_l, \Sigma_l)$ from Gaussian posterior
```

```
Algorithm 8 RLSVI with greedy action
Input: Features $\Phi$; $\lambda > 0$, $\sigma > 0$, $\gamma \in [0,1]$
  $\theta_0^- \leftarrow 0$; $t \leftarrow 0$
  for Episode $l = 0,1,..$ do
      Compute $\tilde{\theta}_l$ using Algorithm 7
      Observe $s_t$
      while TRUE do
          Update $t \leftarrow t+1$
          Sample $a_t \in \text{argmax}_{\alpha \in \mathcal{A}} \left(\Phi \tilde{\theta}\right)(s_t,\alpha)$
          Observe $r_t$ and $s_{t+1}$
          if $s_{t+1}$ is terminal then
              BREAK
          end if
      end while
  end for
```

This algorithm simply approximates a time-homogenous value function using Bayesian linear regression. We found that a discount rate of $\gamma = 0.99$ was helpful for stability in both RLSVI and LSVI.

In order to avoid growing computational and memory cost as LSVI collects more data we used a very simple strategy to only store the most recent N transitions. For our experiments we set $N = 10^5$. Computation for RLSVI and LSVI remained negligible compared to the cost of running the Tetris simulator for our implementations.

To see how small this memory requirement is note that, apart from the number of holes, every feature and reward is a positive integer between 0 and 20 inclusive. The number of holes is a positive integer between 0 and 199. We could store the information 10^5 transitions for every possible action using less than 10mb of memory.

D.2. Effective improvements

We present the results for RLSVI with fixed $\sigma = 1$ and $\lambda = 1$. This corresponds to a Bayesian linear regression with a known noise variance in Algorithm 7. We actually found slightly better performance using a Bayesian linear regression with an inverse gamma prior over an unknown variance. This is the conjugate prior for Gaussian regression with known variance. Since the improvements were minor and it slightly complicates the algorithm we omit these results. However, we believe that using a wider prior over the variance will be more robust in application, rather than picking a specific σ and λ.

D.3. Mini-tetris

In Figure 7 we show that RLSVI outperforms LSVI even with a highly tuned annealing scheme forϵ. However, these results are much more extreme on a didactic version of mini-tetris. We make a tetris board with only 4 rows and only S, Z pieces. This problem is much more difficult and highlights the need for efficient exploration in a more extreme way.

In Figure 14 we present the results for this mini-tetris environment. As expected, this example highlights the benefits of RLSVI over LSVI with dithering. RLSVI greatly outperforms LSVI even with a tuned ϵ schedule. RLSVI learns faster and reaches a higher convergent policy.

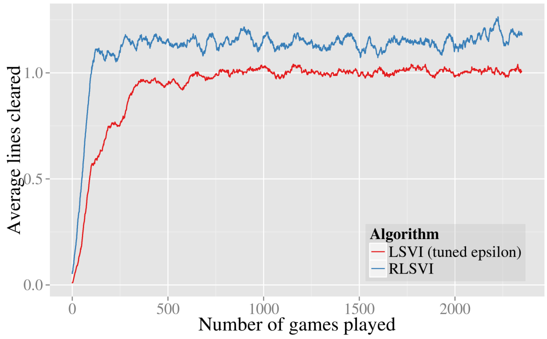

Figure 14. Reduced 4-row tetris with only S and Z pieces.

E. Recommendation system experiments

E.1. Experiment Setup

For the recommendation system experiments, the experiment setup is specified in Algorithm 9. We set $N = 10$, $J = H = 5$, $c = 2$ and $L = 1200$.

Algorithm 9 Recommendation System Experiments: Experiment Setup

```
Input: $N \in \mathbb{Z}_{++}$, $J = H \in \mathbb{Z}_{++}$, $c > 0$, $L \in \mathbb{Z}_{++}$
Output: $\hat{\Delta}(0), \dots, \hat{\Delta}(L-1)$
  for $i = 1, \dots, 100$ do
    Sample a problem instance $\gamma_{an} \sim N(0, c^2)$
    Run the Bernoulli bandit algorithm 100 times
    Run the linear contextual bandit algorithm 100 times
    for for each $\eta \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ do
      Run LSVI-Boltzmann with $\lambda = 0.2$ and $\eta$ 10 times
    end for
    Run RLSVI with $\lambda = 0.2$ and $\sigma^2 = 10^{-3}$ 10 times
  end for
  Compute the average regret for each algorithm
```

The myopic policy is defined as follows: for all episode $l = 0, 1, \cdots$ and for all step $h = 0, \cdots, H-1$, choose $a_{lh} \in \arg\max_a \mathbb{P}(a | x_{lh})$, where a_{lh} and x_{lh} are respectively the action and the state at step h of episode l.

E.2. Bernoulli bandit algorithm

The Bernoulli bandit algorithm is described in Algorithm 10, which is a Thompson sampling algorithm with uniform prior. Obviously, this algorithm aims to learn the myopic policy.

Algorithm 10 Bernoulli bandit algorithm

```
Input: $N \in \mathbb{N}$, $J \in \mathbb{N}$, $L \in \mathbb{N}$
  Initialization: Set $\alpha_n = \beta_n = 1$, $\forall n = 1, 2, \dots, N$
  for $l = 0, \dots, L-1$ do
    Randomly sample $\hat{p}_{ln} \sim \text{beta}(\alpha_n, \beta_n)$, $\forall n = 1, \dots, N$
    Sort $\hat{p}_{ln}$'s in the descending order, and recommend the first $J$ products in order to the customer
    for $n = 1, \dots, N$ do
      if product $n$ is recommended in episode $l$ then
        if customer likes product then
          $\alpha_n \leftarrow \alpha_n + 1$
        else
          $\beta_n \leftarrow \beta_n + 1$
        end if
      end if
    end for
  end for
```

E.3. Linear contextual bandit algorithm

In this subsection, we describe the linear contextual bandit algorithm. The linear contextual bandit algorithm is similar to RLSVI, but without *backward value propagation*, a key feature of RLSVI. It is straightforward to see that the linear contextual bandit algorithm aims to learn the myopic policy. This algorithm is specified in Algorithm 11 and 12. Notice that this algorithm can be implemented incrementally, hence, it is computationally efficient. In this computational study, we use the same basis functions as RLSVI, and the same algorithm parameters (i.e. $\lambda = 0.2$ and $\sigma^2 = 10^{-3}$).

Algorithm 11 Randomized exploration in linear contextual bandits

```
Input: Data $\Phi(s_{i0}, a_{i0}), r_{i0}, .., \Phi(s_{iH-1}, a_{iH-1}), r_{iH} : i < L$
       Parameters $\lambda > 0$, $\sigma > 0$
Output: $\hat{\theta}_{l0}, \dots, \hat{\theta}_{l,H-1}$
  $\hat{\theta}_{lH} \leftarrow 0$, $\Phi_H \leftarrow 0$
  for $h = H-1, \dots, 1, 0$ do
    Generate regression matrix and vector
      $A \leftarrow \begin{bmatrix} \Phi_h(s_{0h}, a_{0h}) \\ \vdots \\ \Phi_h(s_{l-1,h}, a_{l-1,h}) \end{bmatrix}$
      $b \leftarrow \begin{bmatrix} r_{0,h} \\ \vdots \\ r_{l-1,h} \end{bmatrix}$
    Estimate value function
      $\overline{\theta}_{lh} \leftarrow \frac{1}{\sigma^2} \left( \frac{1}{\sigma^2} A^\top A + \lambda \sigma^2 I \right)^{-1} A^\top b$
      $\Sigma_{lh} \leftarrow \left( \frac{1}{\sigma^2} A^\top A + \lambda I \right)^{-1}$
    Sample $\hat{\theta}_{lh} \sim N(\overline{\theta}_{lh}, \Sigma_{lh})$
  end for
```

Algorithm 12 Linear contextual bandit algorithm

```
Input: Features $\Phi_0, .., \Phi_H$, $\sigma > 0$, $\lambda > 0$
  for $l = 0, 1, \cdots$ do
    Compute $\hat{\theta}_{l0}, \dots, \hat{\theta}_{l,H-1}$ using Algorithm 11
    Observe $x_{l0}$
    for $h = 0, \cdots, H-1$ do
      Sample $a_{lh} \sim \text{unif}\left( \text{argmax}_{\alpha \in \mathcal{A}} \left( \Phi_h \hat{\theta}_{lh} \right)(x_{lh}, \alpha) \right)$
      Observe $r_{lh}$ and $x_{l,h+1}$
    end for
  end for
```

F. Extensions

We now briefly discuss a couple possible extensions of the version of RLSVI proposed in Algorithm 1 and 8. One is an incremental version which is computationally more efficient. The other addresses continual learning in an infinite horizon discounted Markov decision process. In the same sense that RLSVI shares much with LSVI but is distinguished by its new approach to exploration, these extensions share much with least-squares Q-learning (Lagoudakis et al., 2002).

F.1. Incremental learning

Note that Algorithm 1 is a batch learning algorithm, in the sense that, in each episode l, though Σ_{lh}'s can be computed incrementally, it needs all past observations to compute $\bar{\theta}_{lh}$'s. Thus, its per-episode compute time grows with l, which is undesirable if the algorithm is applied over many episodes.

One way to fix this problem is to derive an incremental RLSVI that updates $\bar{\theta}_{lh}$'s and Σ_{lh}'s using summary statistics of past data and new observations made over the most recent episode. One approach is to do this by computing

$$\begin{aligned} \Sigma_{l+1,h}^{-1} &\leftarrow (1-\nu_l)\Sigma_{lh}^{-1} + \frac{1}{\sigma^2}\Phi_h\left(x_{lh}, a_{lh}\right)^\top \Phi_h\left(x_{lh}, a_{lh}\right) \\ y_{l+1,h} &\leftarrow (1-\nu_l) y_{lh} + \frac{1}{\sigma^2}\left[r_{lh} + \max_{\alpha\in\mathcal{A}}\left(\Phi_{h+1}\tilde{\theta}_{l,h+1}\right)\left(x_{l,h+1}, \alpha\right)\right]\Phi_h\left(x_{lh}, a_{lh}\right)^\top, \end{aligned} \tag{4}$$

and setting $\bar{\theta}_{l+1,h} = \Sigma_{l+1,h}^{-1} y_{l+1,h}$. Note that we sample $\tilde{\theta}_{lh} \sim N(\bar{\theta}_{lh}, \Sigma_{lh})$, and initialize $y_{0h} = 0$, $\Sigma_{0h}^{-1} = \lambda I$, $\forall h$. The step size ν_l controls the influence of past observations on Σ_{lh} and $\bar{\theta}_{lh}$. Once $\tilde{\theta}_{lh}$'s are computed, the actions are chosen based on Algorithm 8. Another approach would be simply to approximate the solution for $\bar{\theta}_{lh}$ numerically via random sampling and stochastic gradient descent similar to other works with non-linear architectures (Mnih, 2015). The per-episode compute time of these incremental algorithms are episode-independent, which allows for deployment at large scale. On the other hand, we expect the batch version of RLSVI to be more data efficient and thus incur lower regret.

F.2. Continual learning

Finally, we propose a version of RLSVI for RL in infinite-horizon time-invariant discounted MDPs. A discounted MDP is identified by a sextuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \gamma, P, R, \pi)$, where $\gamma \in (0,1)$ is the discount factor. $\mathcal{S}, \mathcal{A}, P, R, \pi$ are defined similarly with the finite horizon case. Specifically, in each time $t = 0, 1, \ldots$, if the state is x_t and an action a_t is selected then a subsequent state x_{t+1} is sampled from $P(\cdot|x_t, a_t)$ and a reward r_t is sampled from $R\left(\cdot|x_t, a_t, x_{t+1}\right)$. We also use V^* to denote the optimal state value function, and Q^* to denote the optimal action-contingent value function. Note that V^* and Q^* do not depend on t in this case.

Algorithm 13 Continual RLSVI

```
Input: $\tilde{\theta}_t \in \mathbb{R}^K$, $w_t \in \mathbb{R}^K$, $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times K}$, $\sigma > 0$, $\lambda > 0$, $\gamma \in (0,1)$, $\{(x_\tau, a_\tau, r_\tau) : \tau \le t\}$, $x_{t+1}$
Output: $\tilde{\theta}_{t+1} \in \mathbb{R}^K$, $w_{t+1} \in \mathbb{R}^K$
1: Generate regression matrix and vector
    $$A \leftarrow \begin{bmatrix} \Phi(x_0, a_0) \\ \vdots \\ \Phi(x_t, a_t) \end{bmatrix} \qquad b \leftarrow \begin{bmatrix} r_0 + \gamma \max_{\alpha\in\mathcal{A}}\left(\Phi\tilde{\theta}_t\right)(x_1, \alpha) \\ \vdots \\ r_t + \gamma \max_{\alpha\in\mathcal{A}}\left(\Phi\tilde{\theta}_t\right)(x_{t+1}, \alpha) \end{bmatrix}$$
2: Estimate value function
    $$\bar{\theta}_{t+1} \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}A^\top A + \lambda I\right)^{-1} A^\top b \qquad \Sigma_{t+1} \leftarrow \left(\frac{1}{\sigma^2}A^\top A + \lambda I\right)^{-1}$$
3: Sample $w_{t+1} \sim N(\gamma w_t, (1-\gamma^2)\Sigma_{t+1})$
4: Set $\tilde{\theta}_{t+1} = \bar{\theta}_{t+1} + w_{t+1}$
```

Similarly with the episodic case, an RL algorithm generates each action a_t based on observations made up to time t, including all states, actions, and rewards observed in previous time steps, as well as the state space $\mathcal{S}$, action space $\mathcal{A}$, discount factor γ, and possible prior information. We consider a scenario in which the agent has prior knowledge that Q^* lies within a linear space spanned by a generalization matrix $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times K}$.

A version of RLSVI for continual learning is presented in Algorithm 13. Note that $\tilde{\theta}_t$ and w_t are values computed by the algorithm in the previous time period. We initialize $\tilde{\theta}_0 = 0$ and $w_0 = 0$. Similarly to Algorithm 1, Algorithm 13 randomly perturbs value estimates in directions of significant uncertainty to incentivize exploration. Note that the random perturbation vectors $w_{t+1} \sim N(\sqrt{1-\gamma^2} w_t, \gamma^2 \Sigma_{t+1})$ are sampled to ensure autocorrelation and that marginal covariance matrices of consecutive perturbations differ only slightly. In each period t, once $\tilde{\theta}_t$ is computed, a greedy action is selected. Avoiding frequent abrupt changes in the perturbation vector is important as this allows the agent to execute on multi-period plans to reach poorly understood state-action pairs.

G. Gaussian vs Dirichlet optimism

The goal of this subsection is to prove Lemma 1, reproduced below:

For all $v \in [0,1]^N$ and $\alpha \in [1,\infty)^N$ with $\alpha^T \mathbb{1} \geq 2$, if $x \sim N(\alpha^\top v / \alpha^\top \mathbf{1}, 1/\alpha^\top \mathbf{1})$ and $y = p^T v$ for $p \sim \text{Dirichlet}(\alpha)$ then $x \succcurlyeq_{\text{so}} y$.

We begin with a lemma recapping some basic equivalences of stochastic optimism.

Lemma 3 (Optimism equivalence)**.**
The following are equivalent to $X \succcurlyeq_{\text{so}} Y$:

1. *For any random variable Z independent of X and Y, $\mathbb{E}[\max(X,Z)] \geq \mathbb{E}[\max(Y,Z)]$*
2. *For any $\alpha \in \mathbb{R}$, $\int_\alpha^\infty \{\mathbb{P}(X \geq s) - \mathbb{P}(Y \geq s)\}\, ds \geq 0$.*
3. *$X =_D Y + A + W$ for $A \geq 0$ and $\mathbb{E}\left[W | Y + A\right] = 0$ for all values $y + a$.*
4. *For any $u : \mathbb{R} \to \mathbb{R}$ convex and increasing $\mathbb{E}[u(X)] \geq \mathbb{E}[u(Y)]$*

These properties are well known from the theory of second order stochastic dominance (Levy, 1992; Hadar & Russell, 1969) but can be re-derived using only elementary integration by parts. $X \succcurlyeq_{\text{so}} Y$ if and only if $-Y$ is second order stochastic dominant for $-X$.

G.1. Beta vs. Dirichlet

In order to prove Lemma 1 we will first prove an intermediate result that shows a particular Beta distribution $\tilde{y}$ is optimistic for y. Before we can prove this result we first state a more basic result that we will use on Gamma distributions.

Lemma 4. *For independent random variables $\gamma_1 \sim Gamma(k_1, \theta)$ and $\gamma_2 \sim Gamma(k_2, \theta)$,*

$$\mathbb{E}[\gamma_1 | \gamma_1 + \gamma_2] = \frac{k_1}{k_1 + k_2}(\gamma_1 + \gamma_2) \qquad \textit{and} \qquad \mathbb{E}[\gamma_2 | \gamma_1 + \gamma_2] = \frac{k_2}{k_1 + k_2}(\gamma_1 + \gamma_2).$$

We can now present our optimistic lemma for Beta versus Dirichlet.

Lemma 5. *Let $y = p^\top v$ for some random variable $p \sim Dirichlet(\alpha)$ and constants $v \in \Re^d$ and $\alpha \in \mathbb{N}^d$. Without loss of generality, assume $v_1 \leq v_2 \leq \cdots \leq v_d$. Let $\tilde{\alpha} = \sum_{i=1}^d \alpha_i (v_i - v_1)/(v_d - v_1)$ and $\tilde{\beta} = \sum_{i=1}^d \alpha_i (v_d - v_i)/(v_d - v_1)$. Then, there exists a random variable $\tilde{p} \sim Beta(\tilde{\alpha}, \tilde{\beta})$ such that, for $\tilde{y} = \tilde{p} v_d + (1 - \tilde{p}) v_1$, $\mathbb{E}[\tilde{y} | y] = \mathbb{E}[y]$.*

Proof. Let $\gamma_i = \text{Gamma}(\alpha, 1)$, with $\gamma_1, \ldots, \gamma_d$ independent, and let $\overline{\gamma} = \sum_{i=1}^d \gamma_i$, so that

$$p \equiv_D \gamma / \overline{\gamma}.$$

Let $\alpha_i^0 = \alpha_i (v_i - v_1)/(v_d - v_1)$ and $\alpha_i^1 = \alpha_i (v_d - v_i)/(v_d - v_1)$ so that

$$\alpha = \alpha^0 + \alpha^1.$$

Define independent random variables $\gamma^0 \sim \text{Gamma}(\alpha_i^0, 1)$ and $\gamma^1 \sim \text{Gamma}(\alpha_i^1, 1)$ so that

$$\gamma \equiv_D \gamma^0 + \gamma^1.$$

Take γ^0 and γ^1 to be independent, and couple these variables with γ so that $\gamma = \gamma^0 + \gamma^1$. Note that $\tilde{\beta} = \sum_{i=1}^d \alpha_i^0$ and $\tilde{\alpha} = \sum_{i=1}^d \alpha_i^1$. Let $\overline{\gamma}^0 = \sum_{i=1}^d \gamma_i^0$ and $\overline{\gamma}^1 = \sum_{i=1}^d \gamma_i^1$, so that

$$1 - \tilde{p} \equiv_D \overline{\gamma}^0 / \overline{\gamma} \qquad \text{and} \qquad \tilde{p} \equiv_D \overline{\gamma}^1 / \overline{\gamma}.$$

Couple these variables so that $1 - \tilde{p} = \overline{\gamma}^0 / \overline{\gamma} \qquad$ and $\qquad \tilde{p} = \overline{\gamma}^1 / \overline{\gamma}$. We then have

$$\begin{aligned}
\mathbb{E}[\tilde{y}|y] &= \mathbb{E}[(1-\tilde{p})v_1 + \tilde{p}v_d | y] = \mathbb{E}\left[\frac{v_1 \overline{\gamma}^0}{\overline{\gamma}} + \frac{v_d \overline{\gamma}^1}{\overline{\gamma}} \Big| y\right] = \mathbb{E}\left[\mathbb{E}\left[\frac{v_1 \overline{\gamma}^0 + v_d \overline{\gamma}^1}{\overline{\gamma}} \Big| \gamma, y\right] \Big| y\right] \\
&= \mathbb{E}\left[\frac{v_1 \mathbb{E}[\overline{\gamma}^0 | \gamma] + v_d \mathbb{E}[\overline{\gamma}^1 | \gamma]}{\overline{\gamma}} \Big| y\right] = \mathbb{E}\left[\frac{v_1 \sum_{i=1}^d \mathbb{E}[\gamma_i^0 | \gamma_i] + v_d \sum_{i=1}^d \mathbb{E}[\gamma_i^1 | \gamma_i]}{\overline{\gamma}} \Big| y\right] \\
&\overset{\text{(a)}}{=} \mathbb{E}\left[\frac{v_1 \sum_{i=1}^d \gamma_i \alpha_i^0 / \alpha_i + v_d \sum_{i=1}^d \gamma_i \alpha_i^1 / \alpha_i}{\overline{\gamma}} \Big| y\right] \\
&= \mathbb{E}\left[\frac{v_1 \sum_{i=1}^d \gamma_i (v_i - v_1) + v_d \sum_{i=1}^d \gamma_i (v_d - v_i)}{\overline{\gamma}(v_d - v_1)} \Big| y\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^d \gamma_i v_i}{\overline{\gamma}} \Big| y\right] = \mathbb{E}\left[\sum_{i=1}^d p_i v_i \Big| y\right] = y,
\end{aligned}$$

where (a) follows from Lemma 4. □

G.2. Gaussian vs Beta

In the previous section we showed that a matched Beta distribution $\tilde{y}$ would be optimistic for the Dirichlet y. We will now show that the Normal random variable x is optimistic for $\tilde{y}$ and so complete the proof of Lemma 1, $x \succcurlyeq_{\text{so}} \tilde{y} \succcurlyeq_{\text{so}} y$.

Unfortunately, unlike the case of Beta vs Dirichlet it is quite difficult to show this optimism relationship between Gaussian x and Beta $\tilde{y}$ directly. Instead we make an appeal to the stronger dominance relationship of single-crossing CDFs.

Definition 2 (Single crossing dominance)**.**
Let X and Y be real-valued random variables with CDFs F_X and F_Y respectively. We say that X single-crossing dominates Y if $\mathbb{E}[X] \geq \mathbb{E}[Y]$ and there a crossing point $a \in \mathbb{R}$ such that:

$$F_X(s) \geq F_Y(s) \iff s \leq a. \tag{5}$$

Note that single crossing dominance implies stochastic optimism. The remainder of this section is devoted to proving that the following lemma:

Lemma 6. *Let $\tilde{y} \sim Beta(\alpha, \beta)$ for any $\alpha > 0, \beta > 0$ and $x \sim N\left(\mu = \frac{\alpha}{\alpha+\beta}, \sigma^2 = \frac{1}{\alpha+\beta}\right)$. Then, x single crossing dominates $\tilde{y}$.*

Trivially, these two distributions will always have equal means so it is enough to show that their CDFs can cross at most once on $(0, 1)$.

G.3. Double crossing PDFs

By repeated application of the mean value theorem, if we want to prove that the CDFs cross at most once on $(0, 1)$ then it is sufficient to prove that the PDFs cross at most twice on the same interval. Our strategy will be to show via mechanical calculus that for the known densities of x and $\tilde{y}$ the PDFs cross at most twice on $(0, 1)$. We lament that the proof as it stands is so laborious, but our attempts at a more elegant solution has so far been unsucessful. The remainder of this appendix is devoted to proving this "double-crossing" property via manipulation of the PDFs for different values of α, β.

We write f_N for the density of the Normal x and f_B for the density of the Beta $\tilde{y}$ respectively. We know that at the boundary $f_N(0-) > f_B(0-)$ and $f_N(1+) > f_B(1+)$ where the ± represents the left and right limits respectively. Since the densities are postive over the interval, we can consider the log PDFs instead.

$$l_B(x) = (\alpha - 1)\log(x) + (\beta - 1)\log(1 - x) + K_B$$

$$l_N(x) = -\frac{1}{2}(\alpha+\beta)\left(x - \frac{\alpha}{\alpha+\beta}\right)^2 + K_N$$

Since $\log(x)$ is injective and increasing, if we could show that $l_N(x) - l_B(x) = 0$ has at most two solutions on the interval we would be done.

Instead we will attempt to prove an even stronger condition, that $l'_N(x) - l'_B(x) = 0$ has at most one solution in the interval. This is not necessary for what we actually want to show, but it is sufficient and easier to deal with since we can ignore the annoying constants.

$$l'_B(x) = \frac{\alpha-1}{x} - \frac{\beta-1}{1-x}$$

$$l'_N(x) = \alpha - (\alpha+\beta)x$$

Finally we will consider an even stronger condition, if $l''_N(x) - l''_B(x) = 0$ has no solution then $l'_B(x) - l'_N(x)$ must be monotone over the region and so it can have at most one root.

$$l''_B(x) = -\frac{\alpha-1}{x^2} - \frac{\beta-1}{(1-x)^2}$$

$$l''_N(x) = -(\alpha+\beta)$$

So now let us define:

$$h(x) := l''_N(x) - l''_B(x) = \frac{\alpha-1}{x^2} + \frac{\beta-1}{(1-x)^2} - (\alpha+\beta) \tag{6}$$

Our goal now is to show that $h(x) = 0$ does not have any solutions for $x \in [0,1]$.

Once again, we will look at the derivatives and analyse them for different values of $\alpha, \beta > 0$.

$$h'(x) = -2\left(\frac{\alpha-1}{x^3} - \frac{\beta-1}{(1-x)^3}\right)$$

$$h''(x) = 6\left(\frac{\alpha-1}{x^4} + \frac{\beta-1}{(1-x)^4}\right)$$

G.3.1. SPECIAL CASE $\alpha > 1$, $\beta \leq 1$

In this region we want to show that actually $g(x) = l'_N(x) - l'_B(x)$ has no solutions. We follow a very similar line of argument and write $A = \alpha - 1 > 0$ and $B = \beta - 1 \leq 0$ as before.

$$g(x) = \alpha - (\alpha+\beta)x + \frac{\beta-1}{1-x} - \frac{\alpha-1}{x}$$

$$g'(x) = h(x) = \frac{A}{x^2} + \frac{B}{(1-x)^2} - (\alpha+\beta)$$

$$g''(x) = h'(x) = -2\left(\frac{A}{x^3} - \frac{B}{(1-x)^3}\right)$$

Now since $B \leq 0$ we note that $g''(x) \leq 0$ and so $g(x)$ is a concave function. If we can show that the maximum of g lies below 0 then we know that there can be no roots.

We now attempt to solve $g'(x) = 0$:

$$\begin{aligned} g'(x) &= \frac{A}{x^2} + \frac{B}{(1-x)^2} = 0 \\ \implies -A/B &= \left(\frac{x}{1-x}\right)^2 \\ \implies x &= \frac{K}{1+K} \in (0,1) \end{aligned}$$

Where here we write $K = \sqrt{-A/B} > 0$. We're ignoring the case of $B = 0$ as this is even easier to show separately. We now evaluate the function g at its minimum $x_K = \frac{K}{1+K}$ and write $C = -B \geq 0$.

$$\begin{aligned}
g(x_K) &= (A+1) - (A+B+2)\frac{K}{1+K} + B(1+K) - A\frac{1+K}{K} \\
&= -AK^2 - AK - A + BK^3 + BK^2 + BK - K^2 + K \\
&= -AK^2 - AK - A - CK^3 - CK^2 - CK - K^2 + K \\
&= -A(A/C) - A(A/C)^{1/2} - A - C(A/C)^{3/2} - C(A/C) - C(A/C)^{1/2} - A/C + (A/C)^{1/2} \\
&= -A^2C^{-1} - A^{3/2}C^{-1/2} - A - A^{3/2}C^{-1/2} - A - A^{1/2}C^{1/2} - AC^{-1} + A^{1/2}C^{1/2} \\
&= -A^2C^{-1} - 2A^{3/2}C^{-1/2} - 2A - AC^{-1} \leq 0
\end{aligned}$$

Therefore we are done with this sub proof. The case of $\alpha \leq 1, \beta > 1$ can be dealt with similarly.

G.3.2. Convex function $\alpha > 1, \beta > 1, (\alpha - 1)(\beta - 1) \geq \frac{1}{9}$

In the case of $\alpha, \beta > 1$ we know that $h(x)$ is a convex function on $(0, 1)$. So now if we solve $h'(x^*) = 0$ and $h(x^*) > 0$ then we have proved our statement. We will write $A = \alpha - 1, B = \beta - 1$ for convenience.

We now attempt to solve $h'(x) = 0$

$$\begin{aligned}
h'(x) &= \frac{A}{x^3} - \frac{B}{(1-x)^3} = 0 \\
\implies A/B &= \left(\frac{x}{1-x}\right)^3 \\
\implies x &= \frac{K}{1+K} \in (0, 1)
\end{aligned}$$

Where for convenience we have written $K = (A/B)^{1/3} > 0$. We now evaluate the function h at its minimum $x_K = \frac{K}{1+K}$.

$$\begin{aligned}
h(x_K) &= A\frac{(K+1)^2}{K^2} + B(K+1)^2 - (A+B+2) \\
&= A(2/K + 1/K^2) + B(K^2 + 2K) - 2 \\
&= 3(A^{2/3}B^{1/3} + A^{1/3}B^{2/3}) - 2
\end{aligned}$$

So as long as $h(x_K) > 0$ we have shown that the CDFs are single crossing. We note a simpler characterization of A, B that guarantees this condition:

$$A, B \geq 1/3 \implies AB \geq 1/9 \implies (A^{2/3}B^{1/3} + A^{1/3}B^{2/3}) \geq 2/3$$

And so we have shown that somehow for α, β large enough away from 1 we are OK. Certianly we have proved the result for $\alpha, \beta \geq 4/3$.

G.3.3. Final region $\{\alpha > 1,\ \beta > 1,\ (\alpha - 1)(\beta - 1) \leq \frac{1}{9}\}$

We now produce a final argument that even in this remaining region the two PDFs are at most double crossing. The argument is really no different than before, the only difficulty is that it is not enough to only look at the derivatives of the log likelihoods, we need to use some bound on the normalizing constants to get our bounds. By symmetry in the problem, it will suffice to consider only the case $\alpha > \beta$, the other result follows similarly.

In this region of interest, we know that $\beta \in (1, \frac{4}{3})$ and so we will make use of an upper bound to the normalizing constant of the Beta distribution, the Beta function.

$$\begin{aligned}
B(\alpha, \beta) &= \int_{x=0}^{1} x^{\alpha-1}(1-x)^{\beta-1} dx \\
&\leq \int_{x=0}^{1} x^{\alpha-1} dx = \frac{1}{\alpha} \qquad (7)
\end{aligned}$$

Our thinking is that, because in $\mathcal{B}$ the value of $\beta - 1$ is relatively small, this approximation will not be too bad. Therefore, we can explicitly bound the log likelihood of the Beta distribution:

$$l_B(x) \geq \tilde{l}_B(x) := (\alpha - 1)\log(x) + (\beta - 1)\log(1-x) + \log(\alpha)$$

We will now make use of a calculus argument as in the previous sections of the proof. We want to find two points $x_1 < x_2$ for which $h(x_i) = l_N''(x) - l_B''(x) > 0$. Since $\alpha, \beta > 1$ we know that h is convex and so for all $x \notin [x_1, x_2]$ then $h > 0$. If we can also show that the gap of the Beta over the maximum of the normal log likelihood

$$\text{Gap} : l_B(x_i) - l_N(x_i) \geq f(x_i) := \tilde{l}_B(x_i) - \max_x l_N(x) > 0 \tag{8}$$

is positive then it must mean there are no crossings over the region $[x_1, x_2]$, since $\tilde{l}_B$ is concave and therefore totally above the maximum of l_N over the whole region $[x_1, x_2]$.

Now consider the regions $x \in [0, x_1)$, we know by consideration of the tails that if there is more than one root in this segment then there must be at least three crossings. If there are three crossings, then the second derivative of their difference h must have at least one root on this region. However we know that h is convex, so if we can show that $h(x_i) > 0$ this cannot be possible. We use a similar argument for $x \in (x_2, 1]$. We will now complete this proof by lengthy amounts of calculus.

Let's remind ourselves of the definition:

$$h(x) := l_N''(x) - l_B''(x) = \frac{\alpha - 1}{x^2} + \frac{\beta - 1}{(1-x)^2} - (\alpha + \beta)$$

For ease of notation we will write $A = \alpha - 1, B = \beta - 1$. We note that:

$$h(x) \geq h_1(x) = \frac{A}{x^2} - (A + B + 2), \;\; h(x) \geq h_2(x) = \frac{B}{(1-x)^2} - (A + B + 2)$$

and we solve for $h_1(x_1) = 0, h_2(x_2) = 0$. This means that

$$x_1 = \sqrt{\frac{A}{A+B+2}}, \;\; x_2 = 1 - \sqrt{\frac{B}{A+B+2}}$$

and clearly $h(x_1) > 0, h(x_2) > 0$. Now, if we can show that, for all possible values of A, B in this region $f(x_i) = l_B(x_i) - \max_x l_N(x) > 0$, our proof will be complete.

We will now write $f(x_i) = f_i(A, B)$ to make the dependence on A, B more clear.

$$f_1(A, B) = \log(1 + A) + A\log\left(\sqrt{\frac{A}{A+B+2}}\right) + B\log\left(1 - \sqrt{\frac{A}{A+B+2}}\right) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(A + B + 2)$$

$$f_2(A, B) = \log(1 + A) + A\log\left(1 - \sqrt{\frac{B}{A+B+2}}\right) + B\log\left(\sqrt{\frac{B}{A+B+2}}\right) + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(A + B + 2)$$

We will now show that $\frac{\partial f_i}{\partial B} \leq 0$ for all of the values in our region $A > B > 0$.

$$\begin{aligned}
\frac{\partial f_1}{\partial B} &= -\frac{A}{2(A+B+2)} + \log\left(1 - \sqrt{\frac{A}{A+B+2}}\right) + \frac{B\sqrt{A}}{2(A+B+2)^{3/2}\left(1-\sqrt{\frac{A}{A+B+2}}\right)} - \frac{1}{2(A+B+2)} \\
&= \frac{1}{2(A+B+2)}\left(\frac{B\sqrt{A}}{\sqrt{A+B+2}\left(1-\sqrt{\frac{A}{A+B+2}}\right)} - A - 1\right) + \log\left(1-\sqrt{\frac{A}{A+B+2}}\right) \\
&= \frac{1}{2(A+B+2)}\left(\frac{B\sqrt{A}}{\sqrt{A+B+2}-\sqrt{A}} - A - 1\right) + \log\left(1-\sqrt{\frac{A}{A+B+2}}\right) \\
&\leq \frac{1}{2(A+B+2)}\left(\frac{\sqrt{B}/3}{\sqrt{A+B+2}-\sqrt{A}} - A - 1\right) - \sqrt{\frac{A}{A+B+2}} \\
&\leq \frac{1}{2(A+B+2)}\left(\frac{1}{3}\sqrt{\frac{B}{B+2}} - A - 1\right) - \sqrt{\frac{A}{A+B+2}} \\
&\leq -\frac{A}{2(A+B+2)} - \sqrt{\frac{A}{A+B+2}} \\
&\leq 0
\end{aligned}$$

and similarly,

$$\begin{aligned}
\frac{\partial f_2}{\partial B} &= -A\left(\frac{\sqrt{\frac{B}{A+B+2}}}{2B} + \frac{1}{2(A+B+2)}\right) + \log\left(\sqrt{\frac{B}{A+B+2}}\right) + B\left(\frac{A+2}{2B(A+B+2)}\right) - \frac{1}{2(A+B+2)} \\
&= \frac{1}{2(A+B+2)}\left(A+2-A-1-A\sqrt{\frac{A+B+2}{B}}\right) + \log\left(\sqrt{\frac{B}{A+B+2}}\right) \\
&= \frac{1}{2(A+B+2)}\left(1-A\sqrt{\frac{A+B+2}{B}}\right) + \frac{1}{2}\log\left(\frac{B}{A+B+2}\right)
\end{aligned}$$

Now we can look at each term to observe that $\frac{\partial^2 f_2}{\partial A \partial B} < 0$. Therefore this expression $\frac{\partial f_2}{\partial B}$ is maximized over A for $A = 0$. We now examine this expression:

$$\frac{\partial f_2}{\partial B}\Big|_{A=0} = \frac{1}{2(B+2)} + \frac{1}{2}\log\left(\frac{B}{B+2}\right) \leq \frac{1}{2}\left(\frac{1}{B+2} + \frac{B}{B+2} - 1\right) \leq 0$$

Therefore, the expressions f_i are minimized at at the largest possible $B = \frac{1}{9A}$ for any given A over our region. We will now write $g_i(A) := f_i(A, \frac{1}{9A})$ for this evalutation at the extremal boundary. If we can show that $g_i(A) \geq 0$ for all $A \geq \frac{1}{3}$ and $i = 1, 2$ we will be done.

We will perform a similar argument to show that g_i is monotone increasing, $g_i'(A) \geq 0$ for all $A \geq \frac{1}{3}$.

$$\begin{aligned}
g_1(A) &= \log(1+A) + A\log\left(\sqrt{\frac{A}{A+\frac{1}{9A}+2}}\right) + \frac{1}{9A}\log\left(1-\sqrt{\frac{A}{A+\frac{1}{9A}+2}}\right) \\
&\quad + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(A+\frac{1}{9A}+2) \\
&= \log(1+A) + \frac{A}{2}\log(A) - \frac{1}{2}(1+A)\log(A+\frac{1}{9A}+2) \\
&\quad + \frac{1}{9A}\log\left(1-\sqrt{\frac{A}{A+\frac{1}{9A}+2}}\right) + \frac{1}{2}\log(2\pi)
\end{aligned}$$

Note that the function $p(A) = A + \frac{1}{9A}$ is increasing in A for $A \geq \frac{1}{3}$. We can conservatively bound g from below noting $\frac{1}{9A} \leq 1$ in our region.

$$\begin{aligned} g_1(A) &\geq\ = \log(1+A) + \frac{A}{2}\log(A) - \frac{1}{2}(1+A)\log(A+3) + \frac{1}{9A}\log\left(1 - \sqrt{\frac{A}{A+2}}\right) + \frac{1}{2}\log(2\pi) \\ &\geq\ \log(1+A) + \frac{A}{2}\log(A) - \frac{1}{2}(1+A)\log(A+3) - \frac{1}{9A}\sqrt{A} + \frac{1}{2}\log(2\pi) =: \tilde{g}_1(A) \end{aligned}$$

Now we can use calculus to say that:

$$\begin{aligned} \tilde{g}_1'(A) &= \frac{1}{A+1} + \frac{1}{A+3} + \frac{\log(A)}{2} + \frac{1}{18A^{3/2}} - \frac{1}{2}\log(A+3) \\ &\geq \frac{1}{A+1} + \frac{1}{A+3} + \frac{1}{18A^{3/2}} + \frac{1}{2}\log(\frac{A}{A+3}) \end{aligned}$$

This expression is monotone decreasing in A and with a limit ≥ 0 and so we can say that $\tilde{g}_1(A)$ is monotone increasing. Therefore $g_1(A) \geq \tilde{g}_1(A) \geq \tilde{g}_1(1/3)$ for all A. We can explicitly evaluate this numerically and $\tilde{g}_1(1/3) > 0.01$ so we are done.

The final piece of this proof is to do a similar argument for $g_2(A)$

$$\begin{aligned} g_2(A) &= \log(1+A) + A\log\left(1 - \sqrt{\frac{\frac{1}{9A}}{A + \frac{1}{9A} + 2}}\right) + \frac{1}{9A}\log\left(\sqrt{\frac{\frac{1}{9A}}{A + \frac{1}{9A} + 2}}\right) \\ &\quad + \frac{1}{2}\log(2\pi) - \frac{1}{2}\log(A + \frac{1}{9A} + 2) \\ &= \log(1+A) + A\log\left(1 - \sqrt{\frac{1}{9A^2 + 18A + 1}}\right) + \frac{1}{2}\left(\frac{1}{9A}\log\left(\frac{1}{9A}\right)\right) \\ &\quad - \frac{1}{2}\left(\frac{1}{9A} + 1\right)\log\left(A + \frac{1}{9A} + 2\right) + \frac{1}{2}\log(2\pi) \\ &\geq \log(1+A) + A\left(-\frac{1}{\sqrt{9A^2}}\right) + \frac{1}{2}\left(\frac{1}{9A}\log\left(\frac{1}{9A}\right)\right) - \frac{1}{2}\left(\frac{1}{3} + 1\right)\log\left(A + \frac{1}{3} + 2\right) + \frac{1}{2}\log(2\pi) \\ &\geq \log(1+A) - \frac{1}{3} - \frac{1}{2e} - \frac{2}{3}\log(A + \frac{7}{3}) + \frac{1}{2}\log(2\pi) =: \tilde{g}_2(A) \end{aligned}$$

Now, once again we can see that $\tilde{g}_2$ is monotone increasing:

$$\begin{aligned} \tilde{g}_2'(A) &= \frac{1}{1+A} - \frac{2/3}{A + 7/3} \\ &= \frac{A+5}{(A+1)(3A+7)} \geq 0 \end{aligned}$$

We complete the argument by noting $g_2(A) \geq \tilde{g}_2(A) \geq \tilde{g}_2(1/3) > 0.01$, which concludes our proof of the PDF double crossing in this region.

G.4. Recap

Using the results of the previous sections we complete the proof of Lemma 6 for Gaussian vs Beta dominance for all possible $\alpha, \beta > 0$ such that $\alpha + \beta \geq 1$. Piecing together Lemma 5 with Lemma 6 completes our proof of Lemma 1. We imagine that there is a much more elegant and general proof method available for future work.