Decision-Making under Uncertainty: Efficiently Estimating where Extra Resources are Needed *

Rémi Munos
CMAP, Ecole Polytechnique, 91128 Palaiseau, France
remi.munos@polytechnique.fr

Abstract

It is desirable that a complex decision-making problem in an uncertain world be adequately modeled by a Markov Decision Process (MDP) whose structural representation is adaptively designed by a *parsimonious* resources allocation process. Resources include time and cost of exploration, amount of memory and computational time allowed for the policy or value function representation. Concerned about making the best use of the available resources, we address the problem of efficiently estimating where adding extra resources is highly needed in order to improve the expected performance of the resulting policy. Possible application in reinforcement learning (RL), when real-world exploration is highly costly, concerns the detection of those areas of the state-space that need primarily to be explored in order to improve the policy. Another application concerns approximation of continuous state-space stochastic control problems using adaptive discretization techniques for which highly efficient grid points allocation is mandatory to survive high dimensionality. Maybe surprisingly these two problems can be formulated under a common framework: for a given resource allocation, which defines a belief function over possible MDPs, find where adding new resources (thus decreasing the uncertainty of some parameters -transition probabilities or rewards) will most likely increase the expected performance of the new policy. We use sample techniques to compute the expected loss when following a given policy (for example the optimal policy of the most probable MDP), and are able to estimate the contribution of each parameter's probability distribution function (pdf) to this loss. This allows to quantify the expected reduction of this loss when reducing the uncertainty over each parameter.

1 Introduction

Assume that we model a complex decision-making problem under uncertainty by a finite MDP. Because of the limited resources used, the parameters of the MDP (transition probabilities and rewards) are uncertain: we assume that we only know a belief function over their possible values. If we select the most probable values of the parameters, we can build a MDP and solve it to deduce

*A short version of this paper appears in NIPS proceedings (Munos, 2001) the corresponding optimal policy. However, because of the uncertainty over the true parameters, this policy may not be the one that maximizes the expected cumulative rewards of the true (but partially unknown) decision-making problem. We can nevertheless use sampling techniques to estimate the expected loss of using this policy. Furthermore, if we assume independence of the parameters (considered as random variables), we are able to derive the contribution of the uncertainty over each parameter to this expected loss. As a consequence, we can predict where adding new resources (thus decreasing the uncertainty over some parameters) will decrease mostly this loss, thus will improve the MDP model of the decision-making problem so as to optimize the expected future rewards.

As possible application, in model-free RL we may wish to minimize the amount of real-world exploration (because each experiment is highly costly). Following (Dearden et al., 1999) we can maintain a Dirichlet pdf over the transition probabilities of the corresponding MDP. Then, our algorithm is able to predict in which parts of the state space we should make new experiments, thus decreasing the uncertainty over some parameters (the posterior distribution being less uncertain than the prior), in order to optimize the expected payoff.

Another application concerns the approximation of continuous (or large discrete) state-space control problems using variable resolution grids, that requires an efficient resource allocation process in order to survive the "curse of dimensionality" in high dimensions. For a given grid, because of the interpolation process, the approximate back-up operator introduces a local interpolation error (see (Munos & Moore, 2000)) that may be considered as a random variable (for example in the random grids of (Rust, 1997)). The algorithm introduced in this paper allows to estimate where we should add new grid-points, thus decreasing the uncertainty over the local interpolation error, in order to increase the expected performance of the new grid representation. The main tool developed here is the calculation of the partial derivative of useful global measures (the value function or the loss of using a sub-optimal policy) with respect to each parameter (probabilities and rewards) of a MDP.

2 Description of the problem

We consider a MDP with a finite state-space X and action-space A. A transition from a state x, action a to a next state y occurs with probability $p(y|x,a)$ and the corresponding (deterministic) reward is $r(x,a)$. We introduce the back-up operator $\mathbf{T}_a$ defined, for any function $W : X \to I\!R$, as

$$\mathbf{T}_a W(x) := \gamma \sum_y p(y|x,a) W(y) + r(x,a), \tag{1}$$

(with some discount factor $0 < \gamma < 1$). It is a contraction mapping, thus the dynamic programming (DP) equation $V(x) = \max_{a \in A} \mathbf{T}_a V(x)$ has a unique fixed point V called the *value function*. Let us define the *Q-values* $Q(x,a) = \mathbf{T}_a V(x)$. The optimal policy π^* is the mapping from any state x to the action $\pi^*(x)$ that maximizes the Q-values: $\pi^*(x) = \max_{a \in A} Q(x,a)$.

The **parameters** of the MDP – the probability and the reward functions – are not perfectly known: all we know is a pdf over their possible values. This uncertainty comes from the limited amount of allocated resources for estimating those parameters.

Let us choose a specific policy $\widehat{\pi}$ (for example the optimal policy of the MDP with the most probable parameters). We can estimate the expected loss of using $\widehat{\pi}$ instead of the true (but unknown) optimal policy π^*. Let us write $\mu = \{\mu_j\}$ the set of all parameters (p and r functions) of a MDP. We assume that we know a probability distribution function $pdf(\mu_j)$ over their possible values. For a MDP M^μ defined by its parameters μ, we write $p^\mu(y|x,a)$, $r^\mu(x,a)$, V^μ, Q^μ, and π^μ respectively its transition probabilities, rewards, value function, Q-values, and optimal policy.

2.1 Direct gain optimization

We define the **gain** $J^\mu(x;\pi)$ in the MDP M^μ as the expected sum of discounted rewards obtained starting from state x and using policy π

$$J^\mu(x;\pi) := \mathbf{E}[\sum_k \gamma^k r^\mu(x_k, \pi(x_k))|x_0 = x;\, \pi], \tag{2}$$

where the expectation is taken for sequences of states $x_k \to x_{k+1}$ occurring with probability $p^\mu(x_{k+1}|x_k, \pi^\mu(x_k))$. By definition, the **optimal gain** in M^μ is $V^\mu(x) = J^\mu(x;\pi^\mu)$ which is obtained for the optimal policy π^μ. Let $\widehat{V^\mu}(x) = J^\mu(x;\widehat{\pi})$ be the **approximate gain** obtained for some approximate policy $\widehat{\pi}$ in the *same* MDP M^μ. We define the **loss to occur** $L^\mu(x)$ from x when one uses the approximate policy $\widehat{\pi}$ instead of the optimal one π^μ in M^μ:

$$L^\mu(x) := V^\mu(x) - \widehat{V^\mu}(x). \tag{3}$$

An example of approximate policy $\widehat{\pi}$ would be the optimal policy of the most probable MDP, defined by the most probable parameters $\widehat{p}(y|x,a)$ and $\widehat{r}(x,a)$.

We also consider the problem of maximizing the global gain from a set of initial states chosen according to some probability distribution $P(x)$. Accordingly, we define the **global gain** of a policy π: $J^\mu(\pi) := \sum_x J^\mu(x;\pi)P(x)$ and the **global loss** L^μ of using some approximate policy $\widehat{\pi}$ instead of the optimal one π^μ

$$L^\mu := J^\mu(\pi^\mu) - J^\mu(\widehat{\pi}) = \sum_x L^\mu(x)P(x). \tag{4}$$

Thus, knowing the pdf over all parameters μ we can define the **expected global loss** $L := \mathbf{E}_\mu[L^\mu]$.

Next, we would like to define what is the contribution of each parameter uncertainty to this loss, so we know where we should add new resources (thus reducing some parameters uncertainty) in order to decrease the expected global loss. We would like to estimate, for each parameter μ_j,

$$\mathbf{E}[\delta L \mid \text{ Add } \delta u \text{ units of resource for} \mu_j]. \tag{5}$$

2.2 Partial derivative of the loss

In order to quantify (5) we need to be more explicit about the pdf over μ considered. First, we assume the independence of the parameters μ_j (considered as random variables). Suppose that $pdf(\mu_j) = \mathcal{N}(0, \sigma_j)$ (normal distribution of mean 0 and standard deviation σ_j). We would like to estimate the variation δL of the expected loss L when we make a small change of the uncertainty over μ_j

(consequence of adding new resources), for example when changing the standard deviation of $\delta\sigma_j$ in $pdf(\mu_j)$. At the limit of an infinitesimal variation we obtain the partial derivative $\frac{\partial L}{\partial \sigma_j}$, which when computed for all parameters μ_j, provides the respective contributions of each parameter uncertainty to the global loss.

Another example is when the $pdf(\mu_j)$ is a uniform distribution of support $[-b_j, b_j]$. Then the partial contribution of μ_j's uncertainty to the global loss can be expressed as $\frac{\partial L}{\partial b_j}$. More generally, we can define a finite number of characteristic scalar measurements of the pdf uncertainty (for example the entropy or the moments) and compute the partial derivative of the expected global loss with respect to these coefficients. Finally, knowing the actual resources needed to estimate a parameter μ_j with some uncertainty defined by $pdf(\mu_j)$, we are able to estimate (5).

2.3 Unbiased estimator

We sample N sets of parameters $\{\mu^i\}_{i=1..N}$ from the $pdf(\mu)$, which define N MDPs M^i. For convenience, we use the superscript ^{i} to refer to the i-th MDP sample and the subscript $_j$ for the j-th parameter of a variable. We solve each MDP using standard DP techniques (see (Puterman, 1994)). This expensive computation that can be speed-up in two ways: first, by using the value function and policy computed for the first MDP as initial values for the other MDPs; second, since all MDPs have the same structure, by computing once for all an efficient ordering (using a topological sort, possibly with loops) of the states that will be used for value iteration.

For each MDP, we compute the global loss L^i of using the policy $\widehat{\pi}$ and estimate the expected global loss: $L \simeq \frac{1}{N}\sum_{i=1}^{N} L^i$. In order to estimate the contribution of a parameter uncertainty to L, we need to derive the partial derivative of L with respect to the characteristic coefficients of $pdf(\mu_j)$. In the case where μ_j follows a normal distribution $\mathcal{N}(0, \sigma_j)$, we can write $\mu_j = \sigma_j \xi_j$ where ξ_j follows $\mathcal{N}(0,1)$. The partial derivative of the expected loss L with respect to σ_j is

$$\frac{\partial L}{\partial \sigma_j} = \frac{\partial}{\partial \sigma_j}\mathbf{E}_{\mu\sim\mathcal{N}(0,\sigma)}[L^\mu] = \frac{\partial}{\partial \sigma_j}\mathbf{E}_{\xi\sim\mathcal{N}(0,1)}[L^{\sigma\xi}] = \mathbf{E}_{\xi\sim\mathcal{N}(0,1)}[\frac{\partial L^{\sigma\xi}}{\partial \mu_j}\xi_j] \qquad (6)$$

from which we deduce the unbiased estimator

$$\frac{1}{N}\sum_{i=1}^{N}\frac{\partial L^i}{\partial \mu_j}\frac{\mu_j^i}{\sigma_j}, \qquad (7)$$

where $\frac{\partial L^i}{\partial \mu_j}$ is the partial derivative of the global loss L^i of MDP M^i with respect to the parameter μ_j (considered as a variable). For other distributions, we have results similar to (6) and we can derive analogous estimators (for uniform distributions, we have the same estimator (7) with b_j instead of σ_j).

Remark: there exists other unbiased estimators of $\frac{\partial L}{\partial \sigma_j}$ that do not require the computation of the partial derivative $\frac{\partial L^i}{\partial \mu_j}$ of the loss with respect to the parameters. However these estimators are much worse than (7) (their variance is higher); the reason being that the expected loss L is a smooth function of the parameters (actually we can prove that it is a convex, piecewise-linear function).

Moreover, as it will be described in next sections, these partial derivatives are possible to calculate and cheap to compute.

The remainder of the paper is organized as follow. Section 3 introduces useful tools to derive the partial contribution of each parameter –transition probability and reward– to the value function in a Markov Chain, Section 4 establishes the partial contribution of each parameter to the global loss, allowing to calculate the estimator (7), and Section 5 provides an efficient algorithm. The proofs are given in the appendices.

3 Non-local dependencies

3.1 Influence of a Markov chain

In (Munos & Moore, 1999) we introduced the notion of influence of a Markov Chain as a way to measure value function/rewards correlations between states. Let us consider a set of values V satisfying a Bellman equation

$$V(x) = \gamma \sum_y p(y|x)V(y) + r(x). \tag{8}$$

We define the discounted cumulative k-chained transition probabilities $p_k(y|x)$ as follow

$$\begin{aligned}
p_0(y|x) &:= \mathbb{1}_{x=y} (= 1 \text{ (if } x = y) \text{ or } 0 \text{ (if } x \neq y)), \\
p_1(y|x) &:= \gamma\, p(y|x), \\
p_2(y|x) &:= \sum_w p_1(y|w) p_1(w|x), \\
&\dots \\
p_k(y|x) &:= \sum_w p_1(y|w) p_{k-1}(w|x).
\end{aligned}$$

The **influence** $I(i|j)$ of a state i on another state j is defined by

$$I(y|x) := \sum_{k=0}^{\infty} p_k(y|x). \tag{9}$$

More intuitively, $I(y|x)$ measures the expected discounted number of visits of state y starting from x; it is also the partial derivative of the value function $V(x)$ with respect to the reward $r(y)$. Indeed $V(x)$ can be expressed as a linear combination of the rewards at y weighted by the influence $I(y|x)$

$$V(x) = \sum_y I(y|x) r(y). \tag{10}$$

We may also define the influence of a state y on a function $f : X \to I\!R$

$$I(y|f(\cdot)) := \sum_x I(y|x) f(x),$$

and the influence of a function $f(\cdot)$ on another function $g(\cdot)$: $I(f(\cdot)|g(\cdot)) := \sum_y f(y) I(y|g(\cdot))$. In (Munos & Moore, 1999), we showed that the influence satisfies

$$I(y|x) = \gamma \sum_w p(y|w) I(w|x) + \mathbb{1}_{x=y}, \tag{11}$$

which is a fixed-point equation of a contraction mapping (in 1−norm) thus has a unique solution −the influence− that can be computed by successive iterations. Similarly, the influence $I(y|f(\cdot))$ may be obtained as limit of iterations

$$I(y|f(\cdot)) \leftarrow \gamma \sum_w p(y|w) I(w|f(\cdot)) + f(y).$$

Thus the computation of the influence $I(y|f(\cdot))$ is cheap (equivalent to solving a Markov chain).

3.2 Differential of V

We wish to express the contribution of each parameter – transition probability p and reward r – (considered as variables) to the value function V by defining the sensitivity of V with respect to those parameters. We already know that the partial derivative of $V(x)$ with respect to the reward $r(z)$ is the influence $I(z|x) = \frac{\partial V(x)}{\partial r(z)}$. Now, we prove that the derivative of $V(x)$ w.r.t. a transition probability $p(w|z)$ is $\gamma I(z|x) V(w)$, which allows to express the differential of V as follows (the proof is in Appendix A).

Theorem 1 *The differential of V is:*

$$dV(x) = \sum_z I(z|x)[\gamma \sum_w V(w) dp(w|z) + dr(z)].$$

4 Differential of the loss

For a given MDP M with parameters μ (for notation simplification we do not write the μ superscript in what follows), we want to estimate the loss of using an approximate policy $\widehat{\pi}$ instead of the optimal one π. First, we define the **one-step loss** $l(x)$ at a state x as the difference between the gain obtained by choosing the best action $\pi(x)$, then the optimal policy π, and the gain obtained by choosing action $\widehat{\pi}(x)$, then the optimal policy π (see figure 1)

$$l(x) := Q(x, \pi(x)) - Q(x, \widehat{\pi}(x)). \tag{12}$$

4.1 Decomposition of the one-step loss.

We use (10) to express the Q-values

$$\begin{aligned} Q(x,a) &= \gamma \sum_w p(w|x,a) \sum_y I(y|w) r(y, \pi(y)) + r(x,a) \\ &= r(x,a) + \sum_y q(y|x,a) r(y, \pi(y)), \end{aligned}$$

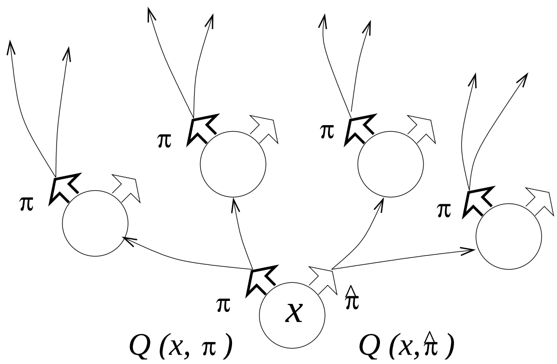

Figure 1: The one-step loss $l(x) = Q(x, \pi(x)) - Q(x, \widehat{\pi}(x))$.

using the partial contributions

$$q(y|x,a) := \gamma \sum_w p(w|x,a) I(y|w), \tag{13}$$

where $I(y|w)$ is the influence of y on w in the Markov chain derived from the MDP M by choosing policy π. Similarly, we decompose the one-step loss

$$\begin{aligned} l(x) &= Q(x,\pi(x)) - Q(x,\widehat{\pi}(x)) \\ &= r(x,\pi(x)) - r(x,\widehat{\pi}(x)) + \sum_y \left[q(y|x,\pi(x)) - q(y|x,\widehat{\pi}(x))\right] r(y,\pi(y)) \\ &= r(x,\pi(x)) - r(x,\widehat{\pi}(x)) + \sum_y l(y|x) r(y,\pi(y)), \end{aligned} \tag{14}$$

as a function of the partial contributions (see figure 2)

$$l(y|x) \quad := \quad q(y|x,\pi(x)) - q(y|x,\widehat{\pi}(x)). \tag{15}$$

4.2 Differential of the one-step loss

Similarly to section (3.2), we express the contribution of the parameters – transition probabilities p and rewards r – (considered as variables) to the one-step loss function by defining the sensitivity of l w.r.t. those parameters.

The differential of the one-step loss is given in the following result (whose proof is in Appendix B).

Theorem 2 *Let us introduce the formal operator* $\mathbf{dT}_a$ *defined, for any function* $W : X \to I\!\!R$, *by*

$$\mathbf{dT}_a W(x) := \gamma \sum_y W(y) dp(y|x,a) + dr(x,a),$$

(similar to the back-up operator (1) but using dp *and* dr *instead of* p *and* r*). The differential of the one-step loss is*

$$dl(x) = \sum_z l(z|x) \mathbf{dT}_{\pi(z)} V(z) + \mathbf{dT}_{\pi(x)} V(x) - \mathbf{dT}_{\widehat{\pi}(x)} V(x).$$

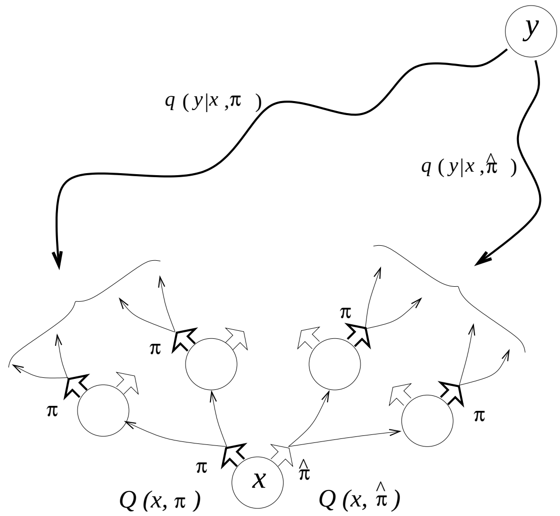

Figure 2: State y contributes to the one-step loss $l(x)$ with the coefficient $l(y|x) = q(y|x, \pi(x)) - q(y|x, \widehat{\pi}(x))$.

4.3 Differential of the global loss

Here, we denote by $\hat{I}(\cdot|\cdot)$ the influence in the Markov chain derived from the MDP M when using the policy $\hat{\pi}$. The proof of this result is given in Appendix C.

Theorem 3 *Let us introduce the operators* $\mathbf{S}$ *and* $\mathbf{dS}$ *defined, for any function* $W : X \to I\!R$*, by*

$$\begin{aligned} \mathbf{S}W(x) &:= \gamma \sum_y W(y) p(y|x, \hat{\pi}(x)) + l(x), \\ \mathbf{dS}W(x) &:= \gamma \sum_y W(y) dp(y|x, \hat{\pi}(x)) + dl(x). \end{aligned}$$

Then, the loss $L(x)$ *at* x *satisfies Bellman's equation* $L = \mathbf{S}L$*. The differential of the loss* $L(x)$ *and global loss* L *are, respectively*

$$\begin{aligned} dL(x) &= \sum_z \hat{I}(z|x) \mathbf{dS}L(z), \\ dL &= \sum_z \hat{I}(z|P(\cdot)) \mathbf{dS}L(z). \end{aligned}$$

After regrouping the contribution to each parameter, we deduce the partial derivatives of the global loss with respect to the rewards and transition probabilities

$$\begin{aligned} \frac{\partial L}{\partial r(x,a)} &= \hat{I}(l(x|\cdot)|P(\cdot)) \mathbb{1}_{a=\pi(x)} + \hat{I}(x|P(\cdot))(\mathbb{1}_{a=\pi(x)} - \mathbb{1}_{a=\widehat{\pi}(x)}), \\ \frac{\partial L}{\partial p(y|x,a)} &= \gamma \hat{I}(x|P(\cdot)) L(y) \mathbb{1}_{a=\hat{\pi}(x)} + \gamma V(y) \frac{\partial L}{\partial r(x,a)}. \end{aligned}$$

5 A sampling algorithm

We use the sampling technique introduced in section 2.3. In order to compute the estimator (7) we calculate the partial derivatives $\frac{\partial L^i}{\partial \mu_j}$ based on the result of the previous section, with the following algorithm.

Given the pdf over the parameters μ, select a policy $\widehat{\pi}$ (for example the optimal policy of the most probable MDP). For $i = 1 \dots N$, solve each MDP M^i and deduce its value function V^i, Q-values Q^i, and optimal policy π^i. Deduce the one-step loss $l^i(x)$ from (12). Compute the influences $\hat{I}(x|P(\cdot))$ and $\hat{I}(l^i(x|\cdot)|P(\cdot))$ from which we deduce $\frac{\partial L^i}{\partial r^i(x,a)}$. Then calculate $L^i(x)$ by solving Bellman's equation $L^i = \mathbf{S}L^i$ and deduce $\frac{\partial L^i}{\partial p^i(y|x,a)}$. These partial derivatives are used in (7) to provide an unbiased estimator of (6).

The complexity of solving a discounted MDP with K states, each one connected to M next states is $\mathcal{O}(KM)$, as is the complexity of computing the influences. Thus, the overall complexity of this algorithm is $\mathcal{O}(NKM)$.

6 Conclusion

Being able to compute the contribution of each parameter –transition probabilities and rewards– to the value function (theorem 1) and to the loss of the expected rewards to occur if we use an approximate policy (theorem 3) enables us to use sampling techniques to estimate what are the parameters whose uncertainty are the most harmful to the expected gain. A relevant resource allocation process would consider adding new computational resources to reduce uncertainty over the true value of those parameters. In the examples given in the introduction, this would be doing new experiments in model-free RL for defining more precisely the transition probabilities of some relevant states. In discretization techniques for continuous control problems, this would be adding new grid points in order to improve the quality of the interpolation at relevant areas of the state space in order to maximize the expected gain of the new policy.

A Proof of theorem 1

A.1 Perturbation of the influence

Consider a perturbation $\delta p(\cdot|z)$ of the probabilities $p(\cdot|z)$. Let us prove that the resulting perturbation of the influence $I(y|x)$ (for any x and y) is

$$\delta I(y|x) = \gamma I(z|x) \sum_w \delta p(w|z) I(y|w). \tag{16}$$

Proof: We first prove by induction on n, that

$$\delta I(y|x) \quad = \quad I(z|x) \left[\sum_{k=0}^{n} \sum_w p_k(y|w) \delta p_1(w|z) \right] + \sum_w p_{n+1}(y|w) \delta I(w|x) \tag{17}$$

Indeed, from (11) we have, for all w,

$$\delta I(w|x) \quad = \quad \sum_{w'} \delta p_1(w|w') I(w'|x) + \sum_{w'} p_1(w|w') \delta I(w'|x)$$

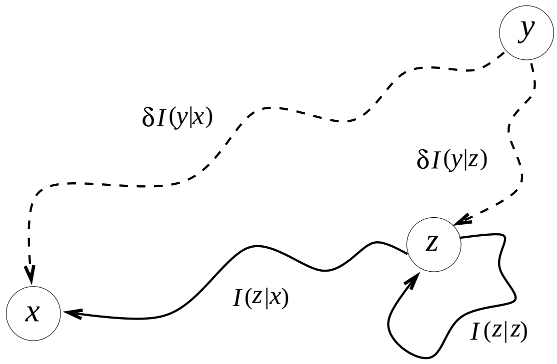

Figure 3: The influence of y on x is the sum of the probabilities of all paths going from x to y. When we perturb a transition probability at z, the paths affected by the change are those going through z. We deduce $\delta I(y|x)I(z|z) = I(z|x)\delta I(y|z)$. The bold arrow represent the influences whereas the dotted arrows show the variation of the influences.

$$= \quad \delta p_1(w|z)I(z|x) + \sum_{w'} p_1(w|w')\delta I(w'|x). \tag{18}$$

For $w = y$ we deduce that (17) is true for $n = 0$. Now assume it is true for n, then from (18),

$$\begin{aligned}
\delta I(y|x) &= I(z|x)[\sum_{k=0}^{n}\sum_{w} p_k(y|w)\delta p_1(w|z) + \sum_{w} p_{n+1}(y|w)p_1(w|z)] \\
&\quad + \sum_{w} p_{n+1}(y|w)\sum_{w'} p_1(w|w')\delta I(w'|x) \\
&= I(z|x)\left[\sum_{k=0}^{n+1}\sum_{w} p_k(y|w)\delta p_1(w|z)\right] + \sum_{w'} p_{n+2}(y|w')\delta I(w'|x),
\end{aligned}$$

and (17) is true for $n+1$, thus for all $n \geq 0$.

Now, when n tends to infinity,

$$\delta I(y|x) = I(z|x)\sum_{k=0}^{\infty}\sum_{w} p_k(y|w)\delta p_1(w|z),$$

which is (16). Q.E.D.

A direct consequence of (16) is that (see figure 3)

$$\delta I(y|x)I(z|z) = I(z|x)\delta I(y|z). \tag{19}$$

A.2 Perturbation of the value function

Consider a perturbation $\delta p(\cdot|z)$ and $\delta r(z)$ of the probabilities $p(\cdot|z)$ and reward $r(z)$ at a state z. The resulting perturbation of the value function at x is

$$\delta V(x) = I(z|x)\left[\gamma\sum_{w}\delta p(w|z)V(w) + \delta r(z)\right]. \tag{20}$$

Proof: From (10) and (19), for all x,

$$\begin{aligned}\delta V(x) &= \sum_y \delta I(y|x)r(y) + \sum_y I(y|x)\delta r(y) \\ &= \frac{I(z|x)}{I(z|z)} \sum_y \delta I(y|z)r(y) + I(z|x)\delta r(z).\end{aligned}$$

Thus for $x = z$,

$$\delta V(z) = \sum_y \delta I(y|z)r(y) + I(z|z)\delta r(z), \tag{21}$$

from which we deduce that

$$\delta V(x) = \frac{I(z|x)}{I(z|z)}\delta V(z). \tag{22}$$

Now, from Bellman's equation,

$$\delta V(z) = \gamma \sum_w \delta p(w|z)V(w) + \gamma \sum_w p(w|z)\delta V(w) + \delta r(z).$$

The middle term of the right hand side of this expression may be rewritten using (22) and (11):

$$\begin{aligned}\gamma \sum_w p(w|z)\delta V(w) &= \gamma \sum_w p(w|z)I(z|w)\frac{\delta V(z)}{I(z|z)} \\ &= [I(z|z) - 1]\frac{\delta V(z)}{I(z|z)}.\end{aligned}$$

Thus:

$$\begin{aligned}\delta V(z) &= \gamma \sum_w \delta p(w|z)V(w) + \delta r(z) + \frac{I(z|z) - 1}{I(z|z)}\delta V(z) \\ &= I(z|z)\left[\gamma \sum_w \delta p(w|z)V(w) + \delta r(z)\right],\end{aligned}$$

which, with (22) proves (20). Q.E.D.

For infinitesimal perturbations at all states z, we deduce the differential of $V(x)$ and theorem 1 follows.

B Proof of theorem 2

Consider a perturbation $\delta p(\cdot|z,a)$ and $\delta r(z,a)$ of the probabilities $p(\cdot|z,a)$ and reward $r(z,a)$ at a state-action z, a.

Assume that $z \neq x$ for now. If a is a sub-optimal action ($a \neq \pi(z)$) then the one-step loss $l(x)$ is left unchanged. Now, consider the case $a = \pi(z)$. Let us write $\delta \mathbf{T}_a V(z)$ the perturbation:

$$\delta \mathbf{T}_a V(z) := \gamma \sum_y \delta p(y|z,a)V(y) + \delta r(z,a).$$

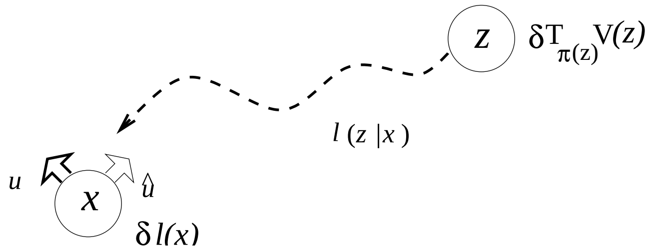

Figure 4: The perturbation of the one-step loss: $\delta l(x) = l(z|x)\delta \mathbf{T}_{\pi(z)} V(z)$.

Then, the resulting perturbation of the one-step loss at x is (see figure 4)

$$\delta l(x) = l(z|x)\delta \mathbf{T}_{\pi(z)} V(z). \tag{23}$$

Proof. From (14),

$$\delta l(x) = \sum_y \delta l(y|x) r(y, \pi(y)) + \sum_y l(y|x) \delta r(y, \pi(y)). \tag{24}$$

From (15), $\delta l(y|x) = \delta q(y|x, \pi(x)) - \delta q(y|x, \widehat{\pi}(x))$ and (13) implies that

$$\delta q(y|x,a) \quad = \quad \gamma \sum_w p(w|x,a) \delta I(y|w)$$

(because $\delta p(\cdot|x,a) = 0$ for $x \neq z$). From (19), $\delta I(y|w) = \frac{I(z|w)}{I(z|z)} \delta I(y|z)$. We deduce that the first term of the right hand side of (24) is

$$\begin{aligned} & \sum_y \delta l(y|x) r(y, \pi(y)) \\ = \quad & \gamma \sum_w \left[p(w|x, \pi(x)) - p(w|x, \widehat{\pi}(x)) \right] \frac{I(z|w)}{I(z|z)} \sum_y \delta I(y|z) r(y, \pi(y)) \\ = \quad & \frac{l(z|x)}{I(z|z)} \sum_y \delta I(y|z) r(y, \pi(y)). \end{aligned}$$

The second term of (24) is $\sum_y l(y|x) \delta r(y, \pi(y)) = l(z|x) \delta r(z, \pi(z))$. By adding the two terms,

$$\delta l(x) = \frac{l(z|x)}{I(z|z)} \left[\sum_y \delta I(y|z) r(y, \pi(y)) + I(z|z) \delta r(z, \pi(z)) \right].$$

Finally, from (21) and (20),

$$\delta l(x) = \frac{l(z|x)}{I(z|z)} \delta V(z) = l(z|x) \delta \mathbf{T}_{\pi(z)} V(z).$$

Q.E.D.

Now, if $z = x$, from (13) the contribution $\gamma \sum_w \delta p(w|z,a) I(y|w)$ needs to be taken into account in $\delta q(y|z,a)$, thus from (14), the additional contribution to $\delta l(x)$ is

$$\begin{aligned}
& \gamma \sum_y \sum_w \left[\delta p(w|x,\pi(x)) - \delta p(w|x,\widehat{\pi}(x))\right] I(y|w) r(y,\pi(y)) \\
& +\delta r(x,\pi(x)) - \delta r(x,\widehat{\pi}(x)) \\
= \;& \gamma \sum_w \left[\delta p(w|x,\pi(x)) - \delta p(w|x,\widehat{\pi}(x))\right] V(w) + \delta r(x,\pi(x)) - \delta r(x,\widehat{\pi}(x)) \\
= \;& \delta \mathbf{T}_{\pi(x)} V(x) - \delta \mathbf{T}_{\widehat{\pi}(x)} V(x),
\end{aligned}$$

which combined with the perturbations (23) for all states z, allows to write the differential of the one-step loss in theorem 2. Q.E.D.

C Proof of theorem 3

The loss $L(x)$, defined by (3),

$$\begin{aligned}
L(x) &= V(x) - \widehat{V}(x) \\
&= \mathbf{T}_{\pi(x)} V(x) - \mathbf{T}_{\widehat{\pi}(x)} \widehat{V}(x) \\
&= \mathbf{T}_{\pi(x)} V(x) - \mathbf{T}_{\widehat{\pi}(x)} V(x) + \mathbf{T}_{\widehat{\pi}(x)} V(x) - \mathbf{T}_{\widehat{\pi}(x)} \widehat{V}(x) \\
&= l(x) + \gamma \sum_y p(y|x,\hat{\pi}(x)) L(y),
\end{aligned}$$

satisfies Bellman's equation $L = \mathbf{S}L$. It is therefore the expected discounted cumulative one-step losses $l(x_k)$ for reachable states x_k using $\widehat{\pi}$:

$$L(x) = \mathbf{E}[\sum_{k\geq 0} \gamma^k l(x_k) | x_0 = x; \widehat{\pi}], \tag{25}$$

From section 3.1 –property (10)– we may decompose $L(x)$ using the influence $\hat{I}(y|x)$:

$$L(x) = \sum_y \hat{I}(y|x) l(y). \tag{26}$$

Consider a perturbation $\delta p(\cdot|z,a)$ and $\delta r(z,a)$ of the probabilities $p(\cdot|z,a)$ and reward $r(z,a)$ at a state-action z, a. Then $L(x)$ is perturbed by

$$\delta L(x) = \sum_y \hat{I}(y|x) \delta l(y) + \sum_y \delta \hat{I}(y|x) l(y).$$

From (19) and (16), the second term is

$$\begin{aligned}
\sum_y \delta \hat{I}(y|x) l(y) &= \frac{\hat{I}(z|x)}{\hat{I}(z|z)} \sum_y \delta \hat{I}(y|z) l(y) \\
&= \gamma \hat{I}(z|x) \sum_w \delta p(w|z,\hat{\pi}(z)) \sum_y \hat{I}(y|w) l(y) \\
&= \gamma \hat{I}(z|x) \sum_w \delta p(w|z,\hat{\pi}(z)) L(w).
\end{aligned}$$

The perturbation $\delta L(x)$ is

$$\delta L(x) = \sum_y \hat{I}(y|x)\delta l(y) + \gamma \hat{I}(z|x) \sum_w \delta p(w|z, \hat{\pi}(z)) L(w).$$

Thus, adding the perturbations from all states z, we derive the differential of the loss:

$$\begin{aligned} dL(x) &= \sum_z \hat{I}(z|x) dl(z) + \sum_z \gamma \hat{I}(z|x) \sum_y dp(y|z, \hat{\pi}(z)) L(y) \\ &= \sum_z \hat{I}(z|x) [\gamma \sum_y dp(y|z, \hat{\pi}(z)) L(y) + dl(z)]. \end{aligned}$$

The differential of the global loss is immediately deduced from $L = \sum_x L(x)P(x)$. Q.E.D.

Now, if we develop dL, we have:

$$\begin{aligned} dL &= \sum_z \hat{I}(z|P(\cdot)) \Bigg[\gamma \sum_y L(y) dp(y|z, \hat{\pi}(z)) \\ &\quad + \sum_w l(w|z) \big[\gamma \sum_t V(t) dp(t|w, \pi(w)) + dr(w, \pi(w)) \big] \\ &\quad + \gamma \sum_t V(t) \big[dp(t|z, \pi(z)) - dp(t|z, \hat{\pi}(z)) \big] + dr(z, \pi(z)) - dr(z, \hat{\pi}(z)) \Bigg]. \end{aligned}$$

Thus the partial derivatives of L w.r.t. the rewards and transition probabilities are:

$$\frac{\partial L}{\partial r(x,a)} = \sum_z \hat{I}(z|P(\cdot)) l(x|z) \mathbb{1}_{a=\pi(x)} + \hat{I}(x|P(\cdot))(\mathbb{1}_{a=\pi(x)} - \mathbb{1}_{a=\hat{\pi}(x)}),$$

and

$$\begin{aligned} \frac{\partial L}{\partial p(y|x,a)} &= \gamma \hat{I}(x|P(\cdot)) L(y) \mathbb{1}_{a=\hat{\pi}(x)} + \gamma \sum_z \hat{I}(z|P(\cdot)) l(x|z) V(y) \mathbb{1}_{a=\pi(x)} \\ &\quad + \gamma \hat{I}(x|P(\cdot)) V(y) (\mathbb{1}_{a=\pi(x)} - \mathbb{1}_{a=\hat{\pi}(x)})] \\ &= \gamma \hat{I}(x|P(\cdot)) L(y) \mathbb{1}_{a=\hat{\pi}(x)} + \gamma V(y) \frac{\partial L}{\partial r(x,a)}. \end{aligned}$$

Q.E.D.

References

Dearden, R., Friedman, N., & Andre, D. (1999). Model based bayesian exploration. *Proceeding of Uncertainty in Artificial Intelligence.*

Munos, R. (2001). Efficient resources allocation for markov decision processes. *Advances in Neural Information Processing Systems.*

Munos, R., & Moore, A. (1999). Influence and variance of a markov chain : Application to adaptive discretizations in optimal control. *Proceedings of the 38th IEEE Conference on Decision and Control.*

Munos, R., & Moore, A. W. (2000). Rates of convergence for variable resolution schemes in optimal control. *International Conference on Machine Learning.*

Puterman, M. L. (1994). *Markov decision processes, discrete stochastic dynamic programming.* A Wiley-Interscience Publication.

Rust, J. (1997). *Using randomization to break the curse of dimensionality.* Computational Economics.