Loss-calibrated Monte Carlo Action Selection

Ehsan Abbasnejad, Justin Domke, Scott Sanner

November 22, 2014

1 Loss-calibrated Monte Carlo Importance Sampling

Sampling in many applications of decision theory is expensive and time-consuming. However, since we know ultimately these samples are used to select high-utility actions, we are interested in guiding the sampler to be more efficient for action selection.

Here, we will pick a distribution q to draw samples from, which are in turn used to select the action that maximizes the EU. The estimated optimal action is $\hat{a}_n$ defined as

$$\hat{a}_n = \arg\max_a \quad \hat{\mathcal{U}}_n(a). \tag{1}$$

Since the samples are drawn randomly from q, $\hat{a}_n$ is a random variable and so is its expected utility $\mathcal{U}(\hat{a}_n)$. As such, we use $\mathbb{E}, \mathbb{P}$ and $\mathbb{V}$ henceforth to denote the expectation, probability and variance operators that handle the random variables that we emphasize are determined by q.

As the optimal action $\hat{a}_n$ has the maximum EU, the optimal q has to maximize the expectation of true EU as shown in the following:

Lemma 1. *For an action set $\mathcal{A} = \{a_1, a_2\}$,*

$$\mathbb{E}[\mathcal{U}(\hat{a}_n)] \geq \Big(-\mathbb{V}\left[\hat{\mathcal{U}}_n(a_1) - \hat{\mathcal{U}}_n(a_2)\right] - \Big(\mathcal{U}(a_1) - \mathcal{U}(a_2)\Big)^2 + 2\Big(\mathcal{U}(a_1) - \mathcal{U}(a_2)\Big)\Big)\Big(\mathcal{U}(a_1) + \mathcal{U}(a_2)\Big). \tag{2}$$

Proof. We know $\mathbb{E}[\mathcal{U}(\hat{a}_n)] = \sum_{a \neq a^*} \mathbb{P}[a = \hat{a}_n]\mathcal{U}(\hat{a}_n) = \sum_{a \neq a^*} \mathbb{E}\left[\mathbb{I}\left[\hat{\mathcal{U}}_n(a) > \max_{a' \in \mathcal{A}\setminus\{a\}} \hat{\mathcal{U}}_n(a')\right]\right]\mathcal{U}(a)$ and since $\mathbb{I}[v > 0] \geq 1 - (v-1)^2$, we can decompose this expectation as

$$\mathbb{E}[\mathcal{U}(\hat{a}_n) \geq \mathbb{E}\left[1 - (\hat{\mathcal{U}}_n(a_1) - \hat{\mathcal{U}}_n(a_2) - 1)^2\right]\mathcal{U}(a_1) + \mathbb{E}\left[1 - (\hat{\mathcal{U}}_n(a_2) - \hat{\mathcal{U}}_n(a_1) - 1)^2\right]\mathcal{U}(a_2).$$

Expanding and rearranging the terms we prove the lemma. □

There are few points to notice in this lemma: (a) the only term in the RHS of Equation 2 that depends on q is the variance, (b) there is an inherent connection between maximizing expected utility and minimizing variance of the difference of estimated EUs, and (c) for more than two actions, the terms on the RHS that depend on q, are weighted by their true (but unknown) EU. Unfortunately, we are unable to employ this simple and yet insightful lemma to obtain optimal q for more than two actions. As such, we turn to regret analysis to further construct over the intuition behind Lemma 1 for multiple actions. To this end, we formulate the problem in three steps to find the optimal q:

1. Firstly, we bound the regret. Since this bound has to be as tight as possible, we establish the connection between regret and the probability of non-optimal action selection in Theorem 2. Although intuitive, this theorem suggests that minimizing the probability of selecting the non-optimal action yields minimal regret and ultimately the optimal distribution of choice, namely q.

2. Since calculating the probability of selecting the non-optimal action is intractable to be directly minimized, we derive an upper bound in Theorem 11. The optimal q is the solution to this theorem. As we will see, the upper bound depends on the simple observation that the estimated EU is greater than or equal to the next best action, i.e. for $a = \hat{a}_n$,

$$\hat{\mathcal{U}}_n(a) \ \geq \max_{a' \in \mathcal{A}\setminus\{a\}} \hat{\mathcal{U}}_n(a').$$

3. Once the upper bounds are established, it can be directly minimized as will be shown in Theorem 14 that yields the optimal value of q. As will be discussed this optimal distribution samples the regions with larger utility difference.

1.1 Minimizing regret

To find the optimal estimated action $\hat{a}_n$ with fewer samples, we wish to select q that minimizes the regret. Formally we define this as

$$\min_q \quad \ell(\hat{a}_n) \qquad \text{where } \ell(\hat{a}_n) = \mathbb{E}\left[\mathcal{U}(a^*) - \mathcal{U}(\hat{a}_n)\right]. \quad (3)$$

As the distribution q governs the samples generated, the minimum in Equation 3 tends to happen when q puts a lot of density on areas of $\boldsymbol{\theta}$ where (a) $p(\boldsymbol{\theta})$ is high and (b) the utility at $\boldsymbol{\theta}$ varies greatly for the different actions. Direct minimization of Equation 3 is difficult, hence we use its connection to the probability of selecting a non-optimal action instead. Tightening this bound with respect to q will lead to a practical strategy. The following theorem shows that the regret can be bounded in terms of the expected utility.

Theorem 2 (Regret bounds)**.** *For the optimal action a^* and its estimate $\hat{a}_n$ the regret as defined in Equation 3, is bounded as*

$$\Delta \mathbb{P}\left[a^* \neq \hat{a}_n\right] \leq \ell(\hat{a}_n) \leq \Gamma \mathbb{P}\left[a^* \neq \hat{a}_n\right], \qquad (4)$$

where $\Delta = \mathcal{U}(a^*) - \max_{a' \in \mathcal{A} \setminus \{a^*\}} \mathcal{U}(a')$ and $\Gamma = \mathcal{U}(a^*) - \min_{a' \in \mathcal{A}} \mathcal{U}(a')$.

Proof. We know $\mathbb{E}\left[\mathcal{U}(\hat{a}_n)\right]$ is equal to

$$\begin{aligned}
&\sum_{a \in \mathcal{A}} \mathbb{P}\left[a = \hat{a}_n\right] \mathcal{U}(a) \\
&= \mathbb{P}\left[a^* = \hat{a}_n\right] \mathcal{U}(a^*) + \sum_{a \in \mathcal{A}, a \neq a^*} \mathbb{P}\left[a = a^*\right] \mathcal{U}(a) \\
&\geq \mathbb{P}\left[a^* = \hat{a}_n\right] \mathcal{U}(a^*) + \sum_{a \in \mathcal{A}, a \neq a^*} \mathbb{P}\left[a = \hat{a}_n\right] \big(\min_{a' \in \mathcal{A}} \mathcal{U}(a')\big)
\end{aligned}$$

and since $\mathbb{P}\left[a^* = \hat{a}_n\right] = 1 - \mathbb{P}\left[a^* \neq \hat{a}_n\right]$ and $\sum_{a \in \mathcal{A}, a \neq a^*} \mathbb{P}\left[a = \hat{a}_n\right] = \mathbb{P}\left[a^* \neq \hat{a}_n\right]$, we have $\ell(\hat{a}_n) \leq \Gamma \mathbb{P}\left[a^* \neq \hat{a}_n\right]$. Similarly,

$$\mathbb{E}\left[\mathcal{U}(\hat{a}_n)\right] \leq (1 - \mathbb{P}\left[a^* \neq \hat{a}_n\right]) \mathcal{U}(a^*) + \mathbb{P}\left[a^* \neq \hat{a}_n\right] \big(\max_{a' \in \mathcal{A} \setminus \{a^*\}} \mathcal{U}(a')\big)$$

which leads to $\ell(\hat{a}_n) \geq \Delta \mathbb{P}\left[a^* \neq \hat{a}_n\right]$. □

The bound is very intuitive: minimizing the probability of the estimated optimal action $\hat{a}_n$ being non-optimal will lead to a bound on the regret. Clearly, for two actions we have $\Delta = \Gamma$. Thus, in the two-action case, minimizing the probability of selecting a non-optimal action is *equivalent* to maximizing the expected utility of the selected action. With more actions, these objectives are not equivalent, but we can see that the difference is controlled in terms of Δ and Γ.

1.2 Minimizing the probability of non-optimal action

We now turn to the problem of minimizing $\mathbb{P}\left[a^* \neq \hat{a}_n\right]$. Before doing so though, we first mention few lemmas that are used in proving the main theorem. In particular, Lemma 3 to 6, provide the necessary bounds on the terms involving $\max$ that are hard to handle. In Lemma 7 we start providing bounds on the probability of non-optimal actions. Further details of the proofs are available in the supplement.

Lemma 3. *Assuming utility values are non-negative everywhere, for a given action $a \neq a^*$ we have*

$$\mathcal{U}(a^*) \leq \mathbb{E}\Big[\max_{a' \in \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a')\Big] \leq \sum_{a' \in \mathcal{A} \setminus \{a\}} \mathcal{U}(a').$$

Proof. Considering Jensen's inequality we have $\max_{a' \in \mathcal{A} \setminus \{a\}} \mathbb{E}[\hat{\mathcal{U}}_n(a')] \leq \mathbb{E}[\max_{a' \in \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a')]$ and from the definition $\mathcal{U}(a^*) = \max_{a' \in \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a')$. Also, since for non-negative utilities we have $\max_{a' \in \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a') \leq \sum \hat{\mathcal{U}}_n(a')$ the lemma is proved. □

Lemma 4. *Assuming utility values are non-negative everywhere, for a given action $a \neq a^*$ we have*

$$\mathbb{E}\left[\hat{\mathcal{U}}_n(a) - \max_{a' \in \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a')\right] \leq \mathcal{U}(a) - \mathcal{U}(a^*).$$

Proof. Applying the expectation to each term and considering Lemma 3 we conclude the proof. □

Lemma 5. *The following bound for a given action $a \neq a^*$ and $\mathfrak{u}(a, \boldsymbol{\theta}) \geq 0$ holds:*

$$\left(\mathbb{E}\Big[\hat{\mathcal{U}}_n(a) - \max_{a' \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a')\Big]\right)^2 \leq \Big(\sum_{a' \in \mathcal{A}} \mathcal{U}(a')\Big)^2. \qquad (5)$$

Proof. From Lemma 3 we can expand the first term as

$$\begin{aligned}
&\left(\mathbb{E}\Big[\hat{\mathcal{U}}_n(a) - \max_{a' \mathcal{A} \setminus \{a\}} \hat{\mathcal{U}}_n(a')\Big]\right)^2 \\
&\leq \mathcal{U}(a)^2 - 2\mathcal{U}(a)\mathcal{U}(a^*) + \Big(\sum_{a' \in \mathcal{A} \setminus \{a\}} \mathcal{U}(a')\Big)^2 \qquad (6)
\end{aligned}$$

and we know $\Big(\sum_{a\in\mathcal{A}}\mathcal{U}(a')\Big)^2=\Big(\mathcal{U}(a)+\sum_{a'\in\mathcal{A}\setminus\{a\}}\mathcal{U}(a')\Big)^2$ which means

$$\Big(\sum_{a'\in\mathcal{A}\setminus\{a\}}\mathcal{U}(a')\Big)^2=\Big(\sum_{a'\in\mathcal{A}}\mathcal{U}(a')\Big)^2-\mathcal{U}(a)^2-2\mathcal{U}(a)\big(\sum_{a'\in\mathcal{A}\setminus\{a\}}\mathcal{U}(a')\big). \tag{7}$$

Substituting Equation 6 in Equation 7 indicates that the difference of the first and second expressions in Equation 5 is always non-negative. □

Lemma 6. *The following bound for a given action $a\neq a^*$ and $\mathfrak{u}(a,\boldsymbol{\theta})\geq 0$ holds:*

$$\left(\mathbb{E}\Big[\max_{a'\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')-\hat{\mathcal{U}}_n(a)\Big]\right)^2\geq\Big(\mathcal{U}(a^*)-\mathcal{U}(a)\Big)^2.$$

Proof. Considering both sides in Lemma 4 are positive when multiplied by -1, we can square them. □

In subsequent lemma, we upper bound the indicator function with a smooth and convex upper bound that will be easier to minimize. The use of *surrogate* function for minimizing indicator has also been used in similar problems (see e.g. [1]).

Lemma 7. *For an optimal action a^* and its estimate $\hat{a}_n$ obtained from sampling, we have $\forall t>0$,*

$$\mathbb{P}\left[a^*\neq\hat{a}_n\right]\leq\sum_{a\neq a^*}\mathbb{E}\left[\left(t\left(\hat{\mathcal{U}}_n(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\right)+1\right)^2\right].$$

Proof. Since we know $\mathbb{I}[v>0]\leq(tv+1)^2$, we have

$$\begin{aligned}\mathbb{P}\left[a^*\neq\hat{a}_n\right]&=\sum_{a\neq a^*}\mathbb{P}[a=\hat{a}_n]\\&=\sum_{a\neq a^*}\mathbb{E}\left[\mathbb{I}\left[\hat{\mathcal{U}}_n(a)>\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\right]\right]\\&\leq\sum_{a\neq a^*}\mathbb{E}\left[\left(t\left(\hat{\mathcal{U}}_n(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\right)+1\right)^2\right].\end{aligned}$$

□

Lemma 8. *Assuming utility values are non-negative everywhere, the following bound holds for some $t>0$:*

$$\begin{aligned}\mathbb{P}\left[a^*\neq\hat{a}_n\right]\leq(k-1)&+2t\sum_{a\in\mathcal{A}\setminus\{a^*\}}\Big(\mathcal{U}(a)-\mathcal{U}(a^*)\Big)\\&+t^2\sum_{a\in\mathcal{A}\setminus\{a^*\}}\mathbb{V}\bigg[\Big(\hat{\mathcal{U}}_n(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\Big)\bigg]\\&+t^2\Big(\sum_{a'\in\mathcal{A}}\mathcal{U}(a')\Big)^2.\end{aligned}$$

Proof. Expanding the RHS of Lemma 7 we have

$$\begin{aligned}&\mathbb{E}\left[\left(t\left(\hat{\mathcal{U}}_n(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\right)+1\right)^2\right]\\&=1+2t\mathbb{E}\bigg[\Big(\hat{\mathcal{U}}_n(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\Big)\bigg]\\&\quad+t^2\mathbb{E}\bigg[\Big(\hat{\mathcal{U}}_n(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\hat{\mathcal{U}}_n(a')\Big)^2\bigg].\end{aligned}$$

From Lemma 4 we can expand the first expectation and then considering $\mathbb{E}[X^2]=\mathbb{V}[X]+\mathbb{E}[X]^2$ and Lemma 5 we get the bounds in the Lemma 8. □

Lemma 9. *For the bounds detailed in Lemma 8, the value of t that minimizes the upper bound of RHS is*

$$t\quad=\quad\frac{\Delta}{\Big(\sum_{a\in\mathcal{A}}\mathcal{U}(a)\Big)^2},$$

where $\Delta=\mathcal{U}(a^*)-\max_{a'\in\mathcal{A}\setminus\{a^*\}}\mathcal{U}(a')$.

Proof. We know $\Delta=\mathcal{U}(a^*)-\max_{a'\in\mathcal{A}\setminus\{a^*\}}\mathcal{U}(a')\leq\mathcal{U}(a^*)-\mathcal{U}(a)$ considering $\max_{a'\in\mathcal{A}\setminus\{a^*\}}\mathcal{U}(a')\geq\mathcal{U}(a)$ for $a\neq a^*$, then

$$\mathcal{U}(a)-\max_{a'\in\mathcal{A}\setminus\{a\}}\mathcal{U}(a')=\mathcal{U}(a)-\mathcal{U}(a^*)\quad\leq\quad-\Delta$$

and we can rewrite Lemma 8 by replacing the second term with its upper bound (because the variance decreases with the number of samples n to ultimately approach zero we disregard it here) as:

$$\mathbb{P}\left[a^*\neq\hat{a}_n\right]\leq(k-1)\left(1-2t\Delta+t^2\Big(\sum_{a\in\mathcal{A}}\mathcal{U}(a)\Big)^2\right), \tag{8}$$

taking the derivative of RHS. with respect to t and equating to zero we will get the solution. □

Lemma 10. *The following bounds on $\mathbb{P}[a^*\neq\hat{a}_n]$ holds as $n\to+\infty$:*

$$\mathbb{P}\left[a^*\neq\hat{a}_n\right]\quad\leq\quad(k-1)\left(1-\Big(\frac{\Delta}{\sum_{a\in\mathcal{A}}\mathcal{U}(a)}\Big)^2\right). \tag{9}$$

Proof. Replacing the value of t in the bounds in Equation 8 yields the proof. □

In words, the probability of estimating the optimal action increases in proportion with the gap between the expected utility of the best and second best actions. Also if, without loss of generality, we assume that the minimum value of the expected utility is zero, then for two action case the RHS of Equation 9 is also zero that indicates the bound in this lemma is tight. These bounds are most didactic in two action case.

Putting everything together, the following theorem bounds the probability of non-optimal action selection:

Theorem 11 (Upper bound on the probability of non-optimal actions)**.** *We have the following upper bound probability of non-optimal action selection for k actions in set $\mathcal{A}$, true expected utility $\mathcal{U}(a)$ and its estimation $\hat{\mathcal{U}}_n(a)$ obtained from finite samples:*

$$\mathbb{P}\left[a^* \neq \hat{a}_n\right] \leq (k-1) + \sum_{a \in \mathcal{A} \setminus \{a^*\}} t \Bigg(\Delta + 2\Big(\mathcal{U}(a) - \mathcal{U}(a^*)\Big) + t\mathbb{V}\left[\hat{\mathcal{U}}_n(a) - \max_{a' \in \mathcal{A} \setminus a} \hat{\mathcal{U}}_n(a')\right] \Bigg), \tag{10}$$

where t is given in Lemma 9.

Proof. Replacing the value of t obtained in Lemma 9 in Lemma 8, we have this theorem. □

The critical feature of Equation 10 is that all terms on the RHS other than the variance are constant with respect to the sampling distribution q. Thus, this theorem suggests that a reasonable surrogate to minimize the regret in Equation 3 and consequently maximize the expected utility of the estimated optimal action is to minimize the variance of the difference of the estimated utilities. This result is quite intuitive – if we have a low-variance estimate of the differences of utilities, we will tend to select the best action.

This is aligned with the importance sampling literature where it is well known that the optimal distribution to sample from is the one that minimizes the variance [4, 2] with a closed form solution as summarized in the following:

Corollary 12. *[4] Define*

$$\mathbb{E}[H(\mathbf{s})] = \int H(\mathbf{s}) f(\mathbf{s}) d\mathbf{s} = \int H(\mathbf{s}) \frac{f(\mathbf{s})}{g(\mathbf{s})} g(\mathbf{s}) d\mathbf{s}.$$

Then, the solution to the variance minimization problem

$$\min_{g} \quad \mathbb{V}\left[H(\mathbf{s}) \frac{f(\mathbf{s})}{g(\mathbf{s})}\right]$$

is given by

$$g^*(\mathbf{s}) = \frac{|H(\mathbf{s})|\, f(\mathbf{s})}{\int |H(\mathbf{s})|\, f(\mathbf{s}) d\mathbf{s}}.$$

Our analysis shows the variance of the function that has to be minimized is of a particular form that depends on the difference of the utilities (rather than each utility independently).

Lemma 13. *We have the following bound on the sum of variances*

$$\sum_{a \in \mathcal{A} \setminus \{a^*\}} \mathbb{V}\left[\max_{a' \in \mathcal{A} \setminus a} \hat{\mathcal{U}}_n(a') - \hat{\mathcal{U}}_n(a)\right]\Bigg) \leq \sum_{a \in \mathcal{A} \setminus \{a^*\}} \mathbb{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} \Upsilon(\boldsymbol{\theta}_i, a)\right)^2\right] - C, \tag{11}$$

where $\Upsilon(\boldsymbol{\theta}_i, a) = \frac{p(\boldsymbol{\theta}_i)}{q(\boldsymbol{\theta}_i)} \left(\max_{a' \in \mathcal{A} \setminus \{a\}} \mathfrak{u}(\boldsymbol{\theta}_i, a') - \mathfrak{u}(\boldsymbol{\theta}_i, a)\right)$ and $C = \Big(\mathcal{U}(a^*) - \mathcal{U}(a)\Big)^2$.

Proof. We know $\mathbb{V}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$ and the upper bound on the first term is proved using the Jensen's inequality (the weights are normalized as will be discussed in Equation 13). Also, from Lemma 6 we have the second term. □

1.3 Optimal q

We established that to find the optimal proposal distribution q^* (i.e. optimal q), we minimize the sum of variances obtained from Theorem 11. Since a^* is unknown, we sum over all actions in $\mathcal{A}$, rather than just $\mathcal{A} \setminus \{a^*\}$. Since C is independent of q in Equation 11, the objective is to minimize the RHS subject to $\int q(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$ so that the resulting solution is a proper distribution.

The following theorem provides the solution to the optimization problem in Equation 11 that we are interested in:

Theorem 14. *Let $\mathcal{A} = \{a_1, \ldots, a_k\}$ with non-negative utilities. The optimal distribution $q^*(\boldsymbol{\theta})$ is the solution to problem in Equation 11 and has the following form:*

$$q^*(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta}) \sqrt{\sum_{a \in \mathcal{A}} \left(\max_{a' in \mathcal{A} \setminus \{a\}} \mathfrak{u}(\boldsymbol{\theta}, a') - \mathfrak{u}(\boldsymbol{\theta}, a)\right)^2}. \tag{12}$$

Proof. We have the following value to minimize:

$$\frac{1}{n^2} \int \sum_{i=1}^{n} \sum_{j=1}^{n} \Upsilon(\boldsymbol{\theta}_i, a) \Upsilon(\boldsymbol{\theta}_j, a) q(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) d\boldsymbol{\theta}_1 \ldots \boldsymbol{\theta}_n.$$

Since all the samples are independent, the joint distribution factorizes as follows: $q(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = q(\boldsymbol{\theta}_1) \ldots q(\boldsymbol{\theta}_n)$. Now if $i \neq j$, it is easy to see that q vanishes and those terms become independent of q. If $i = j$ however, we have one of the terms in the denominator canceled out with the joint. Also because the sum is over similar terms, we have n times the same expression that lead to the Lagrangian of the optimization to become:

$$\mathcal{L}(q, \lambda) = \frac{1}{n} \sum_{a \in \mathcal{A}} \int \frac{\Upsilon(\boldsymbol{\theta}, a)^2 p(\boldsymbol{\theta})^2}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} + \lambda \left(\int q(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 \right).$$

Taking the derivative with respect to a fixed $q(\boldsymbol{\theta})$, we have

$$-\frac{1}{n} \sum_{a \in \mathcal{A}} \frac{\Upsilon(\boldsymbol{\theta}, a)^2 p(\boldsymbol{\theta})^2}{q(\boldsymbol{\theta})^2} + \lambda = 0 \Rightarrow \sum_{a \in \mathcal{A}} \frac{p(\boldsymbol{\theta})^2}{q(\boldsymbol{\theta})^2} \Upsilon(\boldsymbol{\theta}, a)^2 = \lambda n$$

which concludes the theorem since λn only induces a proportionality constant. □

This is quite intuitive – the samples $\boldsymbol{\theta}$ will be concentrated on regions where $p(\boldsymbol{\theta})$ is large, and the difference of utilities between the actions is large, which is precisely the intuition that motivated our work in Figure **??**. This will tend to lead to the empirically optimal action being the true one, i.e. that $\hat{a}_n$ approaches a^*.

Since the distributions that are used in practice are commonly unnormalized and their normalizers are often unknown for p and certainly unknown for q, we simply remark that the application of loss-calibrated importance sampling requires the following well-known self-normalized variant of $\mathcal{U}(a)$ from Equation **??** for unnormalized distributions $p(\boldsymbol{\theta}) \propto \tilde{p}(\boldsymbol{\theta})$ and $q(\boldsymbol{\theta}) \propto \tilde{q}(\boldsymbol{\theta})$:

$$\begin{aligned} \mathcal{U}(a) &= \int \mathfrak{u}(\boldsymbol{\theta}, a) \tilde{p}(\boldsymbol{\theta}) d\boldsymbol{\theta} \Big/ \int \tilde{p}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \mathfrak{u}(\boldsymbol{\theta}, a) \frac{\tilde{p}(\boldsymbol{\theta})}{\tilde{q}(\boldsymbol{\theta})} \tilde{q}(\boldsymbol{\theta}) d\boldsymbol{\theta} \Big/ \int \frac{\tilde{p}(\boldsymbol{\theta})}{\tilde{q}(\boldsymbol{\theta})} \tilde{q}(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \frac{1}{n} \sum_{i=1}^{n} \mathfrak{u}(\boldsymbol{\theta}_i, a) \frac{\tilde{p}(\boldsymbol{\theta}_i)}{\tilde{q}(\boldsymbol{\theta}_i)} \Big/ \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{p}(\boldsymbol{\theta}_i)}{\tilde{q}(\boldsymbol{\theta}_i)} \quad \boldsymbol{\theta}_i \sim \tilde{q}. \end{aligned} \tag{13}$$

This simply means that for the case of unnormalized $\tilde{p}$ and $\tilde{q}$, all the utility values have to now be reweighted by the slightly more complex $\left(\frac{\tilde{p}(\boldsymbol{\theta}_i)}{\tilde{q}(\boldsymbol{\theta}_i)} \Big/ \sum_{j=1}^{n} \frac{\tilde{p}(\boldsymbol{\theta}_j)}{\tilde{q}(\boldsymbol{\theta}_j)} \right)$.

Furthermore, as it is hard to directly sample q, we must resort to Markov Chain Monte Carlo (MCMC) methods [3]. One of the more commonly used MCMC methods is Metropolis-Hastings (MH). It works by modeling the selection of every sample by considering the probability of jump from a given sample to a new one based on a proposal distribution (often an isotropic Gaussian in practice).

References

[1] Peter Bartlett, I. Michael Jordan, and Jon D. Mcauliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, March 2006.

[2] Paul Glasserman. *Monte Carlo Methods in Financial Engineering*. Applications of Mathematics. Springer, 1st edition, 2004.

[3] Radford M. Neal. Probabilistic inference using markov chain monte carlo methods. Technical report, University of Toronto, 1993.

[4] Reuven Y. Rubinstein. *Simulation and the Monte Carlo Method*. John Wiley & Sons, Inc., 1st edition, 1981.