Actor-Critic Algorithms for Risk-Sensitive MDPs

Prashanth L.A., Mohammad Ghavamzadeh

Actor-Critic Algorithms for Risk-Sensitive MDPs

Prashanth L.A. PRASHANTH.LA@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL

Mohammad Ghavamzadeh MOHAMMAD.GHAVAMZADEH@INRIA.FR
INRIA Lille - Nord Europe, Team SequeL

Abstract

In many sequential decision-making problems we may want to manage risk by minimizing some measure of variability in rewards in addition to maximizing a standard criterion. Variance related risk measures are among the most common risk-sensitive criteria in finance and operations research. However, optimizing many such criteria is known to be a hard problem. In this paper, we consider both average and discounted reward Markov decision processes. For each formulation, we first define a measure of variability for a policy, which in turn gives us a set of risk-sensitive criteria to optimize. For each of these criterion, we derive a formula for computing its gradient. We then devise actor-critic algorithms for estimating the gradient and updating the policy parameters in the ascent direction. We establish the convergence of our algorithms to locally risk-sensitive optimal policies. Finally, we demonstrate the usefulness of our algorithms in a traffic signal control application.

1. Introduction

The usual optimization criteria for an infinite horizon Markov decision process (MDP) are the *average reward* and *expected sum of discounted rewards*. Many algorithms have been developed to maximize these criteria both when the model of the system is known (planning) and unknown (learning). These algorithms can be categorized to value function based methods that are mainly based on the two celebrated dynamic programming algorithms *value iteration* and *policy iteration*; and policy gradient methods that are based on updating the policy parameters in the direction of the gradient of a performance measure (the average reward or the value function of the initial state). However in many applications, we may prefer to minimize some measure of *risk* as well as maximizing a usual optimization criterion. In such cases, we would like to use a criterion that incorporates a penalty for the *variability* induced by a given policy. This variability can be due to two types of uncertainties: bf 1) uncertainties in the model parameters, which is the topic of *robust* MDPs (e.g., Nilim & Ghaoui 2005; Delage & Mannor 2010; Xu & Mannor 2012), and **2)** the inherent uncertainty related to the stochastic nature of the system, which is the topic of *risk-sensitive* MDPs (e.g., Howard & Matheson 1972).

In risk-sensitive sequential decision-making, the objective is to maximize a risk-sensitive criterion such as the expected exponential utility (Howard & Matheson, 1972), a variance related measure (Sobel, 1982; Filar et al., 1989), or the percentile performance (Filar et al., 1995). The issue of how to construct such criteria in a manner that will be both conceptually meaningful and mathematically tractable is still an open question. Although risk-sensitive sequential decision-making has a long history in operations research and finance, it has only recently grabbed attention in the machine learning community. This is why most of the work on this topic (including those mentioned above) has been in the context of MDPs (when the model is known) and much less work has been done within the reinforcement learning (RL) framework. In risk-sensitive RL, we can mention the work by Borkar (2001; 2002) who considered the expected exponential utility and the one by Tamar et al. (2012) on several variance related measures. Tamar et al. (2012) considered episodic problems and proposed a policy gradient algorithm for maximizing several risk-sensitive criteria that involve both the expectation and variance of the *return* (defined as the sum of rewards received in an episode).

In this paper, we consider both average and discounted reward MDPs and take into account variance-related risk measures. In the average reward formulation, we define the measure of variability as the *long-run variance* of a policy, while in the discounted reward setting we define it as the *variance of the return* (similar to Tamar et al. 2012). These

Proceedings of the 30^{th} International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

variability measures in turn define a set of risk-sensitive criteria for each MDP formulation, among which we mainly focus on **1)** maximizing the mean subject to the variance being bounded from above, and **2)** the Sharpe ratio which is popular in financial decision-making (Sharpe, 1966). For each of these risk-sensitive criterion, we first derive a formula for computing its gradient, and then devise actor-critic algorithms for estimating the gradient (while maintaining an estimate of the value function) and updating the policy parameters in the ascent direction. In our average reward actor-critic algorithm, we estimate the gradient by directly working with the obtained formula. However, this approach cannot be easily extended to the discounted reward setting where the risk-sensitive criterion involves both the expectation and variance of the return (see the discussion in Section 6). Hence in the discounted formulation, we estimate the gradient using two simultaneous perturbation methods: *simultaneous perturbation stochastic approximation* (SPSA) (Spall, 1992) and *smoothed functional* (SF) (Katkovnik & Kulchitsky, 1972), resulting in two separate discounted reward actor-critic algorithms. Finally, we establish the convergence of our algorithms to locally risk-sensitive optimal policies and show their usefulness in a traffic signal control problem.

2. Preliminaries

We consider problems in which the agent's interaction with the environment is modeled as a MDP. A MDP is a tuple $(\mathcal{X}, \mathcal{A}, q, P, P_0)$ where $\mathcal{X} = \{1, \ldots, n\}$ and $\mathcal{A} = \{1, \ldots, m\}$ are the state and action spaces, respectively; $q(\cdot|x,a)$ is the probability distribution over rewards; $P(\cdot|x,a)$ is the transition probability distribution; (we assume that P and q are stationary); and $P_0(\cdot)$ is the initial state distribution. We denote the random variable distributed according to $q(\cdot|x,a)$ by $R(x,a)$ and its expectation by $r(x,a) = \mathbb{E}\big[R(x,a)\big]$. In addition, we need to specify the rule according to which the agent selects actions at each state. We assume that this rule does not depend explicitly on time. A *stationary policy* $\mu(\cdot|x)$ is a probability distribution over actions, conditioned on the current state. In policy gradient and actor-critic methods, we define a class of parameterized stochastic policies $\big\{\mu(\cdot|x;\theta), x \in \mathcal{X}, \theta \in \Theta \subseteq \mathbb{R}^{d_1}\big\}$, estimate the gradient of a performance measure w.r.t. the policy parameters $theta$ from the observed system trajectories, and then improve the policy by adjusting its parameters in the direction of the gradient. Since in this setting a policy μ is represented by its d_1-dimensional parameter vector θ, policy dependent functions can be written as a function of θ in place of μ. The following assumption is a standard requirement in policy gradient and actor-critic methods:

(A1) *For any state-action pair (x,a), $\mu(a|x;\theta)$ is continuously differentiable in the parameter θ.*

3. Average Reward Setting

In an average reward MDP, we assume

(A2) *The Markov chain induced by any policy is irreducible and aperiodic.*

The average reward per step under policy μ is defined as

$$\rho(\mu) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} R_t \mid \mu\right] = \sum_{x,a} d^\mu(x)\mu(a|x)r(x,a),$$

where $d^\mu(x)$ is the stationary distribution of state x under policy μ. This limit is well-defined under (A2). The goal in the standard average reward formulation is to find an *average optimal* policy, i.e., $\mu^* = \arg\max_\mu \rho(\mu)$. In this setting, a policy μ is assessed according to the expected differential reward associated with states or state-action pairs. For all states $x \in \mathcal{X}$ and actions $a \in \mathcal{A}$, the *differential* action-value and value functions of policy μ are defined as

$$Q^\mu(x,a) = \sum_{t=0}^{\infty} \mathbb{E}\big[R_t - \rho(\mu) \mid x_0 = x, a_0 = a, \mu\big],$$
$$V^\mu(x) = \sum_a \mu(a|x) Q^\mu(x,a).$$

These functions satisfy the following Poisson equations (Puterman, 1994)

$$\rho(\mu) + V^\mu(x) = \sum_a \mu(a|x)\big[r(x,a) + \sum_{x'} P(x'|x,a)V^\mu(x')\big],$$
$$\rho(\mu) + Q^\mu(x,a) = r(x,a) + \sum_{x'} P(x'|x,a)V^\mu(x'). \quad (1)$$

Different criteria have been proposed to define a measure of *variability* in the average reward formulation, among which we consider the *long-run variance* of μ (Filar et al., 1989)

$$\Lambda(\mu) = \sum_{x,a} \pi^\mu(x,a)\big[r(x,a) - \rho(\mu)\big]^2 \quad (2)$$
$$= \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1} \big(R_t - \rho(\mu)\big)^2 \mid \mu\right],$$

where $\pi^\mu(x,a)$ is the stationary distribution of state-action pair (x,a) under policy μ. This notion of variability is based on the observation that it is the frequency of occurrence of state-action pairs that determine the variability in the average reward. It is easy to show that

$$\Lambda(\mu) = \eta(\mu) - \rho(\mu)^2, \quad \eta(\mu) = \sum_{x,a} \pi^\mu(x,a) r(x,a)^2.$$

We consider the following risk-sensitive measure for average reward MDPs in this paper:

$$\max_\theta \rho(\theta) \qquad \text{subject to} \qquad \Lambda(\theta) \le \alpha. \quad (3)$$

The goal would be to find a policy θ^* (or μ^*) that maximizes (3). The proposed algorithm can be easily extended to other variance-based risk measures including the Sharpe Ratio (SR), i.e., $\max_\theta S(\theta) = \max_\theta \rho(\theta)/\sqrt{\Lambda(\theta)}$.

4. Average Reward Algorithm

To solve the constrained optimization problem (3), we employ the Lagrangian relaxation procedure (Bertsekas, 1999) to convert it to the unconstrained problem

$$\max_{\lambda}\min_{\theta} L(\theta,\lambda) = \max_{\lambda}\min_{\theta}\big[-\rho(\theta)+\lambda\big(\Lambda(\theta)-\alpha\big)\big],$$

where λ is the Lagrange multiplier associated with the variance constraint $\Lambda(\theta) \leq \alpha$. The goal here is to find the saddle point of $L(\theta,\lambda)$, i.e., a point (θ^*,λ^*) that satisfies $L(\theta,\lambda^*) \geq L(\theta^*,\lambda^*) \geq L(\theta^*,\lambda), \forall\theta, \forall\lambda > 0$.

In order to optimize (3) we should descend in θ and ascend in λ. The descent and ascent can be performed using $\nabla_\theta L(\theta,\lambda) = -\nabla_\theta\rho(\theta) + \lambda\nabla_\theta\Lambda(\theta)$ and $\nabla_\lambda L(\theta,\lambda) = \Lambda(\theta)-\alpha$, respectively.

Since $\nabla\Lambda(\theta) = \nabla\eta(\theta) - 2\rho(\theta)\nabla\rho(\theta)$, in order to compute $\nabla\Lambda(\theta)$ it would be enough to calculate $\nabla\eta(\theta)$. Before stating the expressions for the gradients of $\rho(\theta)$ and $\eta(\theta)$, we define U^μ and W^μ as the differential value and action-value functions associated with the square reward under policy μ. Similar to V^μ and Q^μ (Eq. 1), U^μ and W^μ satisfy the following Poisson equations:

$$\eta(\mu) + U^\mu(x) = \sum_a \mu(a|x)\big[r(x,a)^2 + \sum_{x'} P(x'|x,a)U^\mu(x')\big],$$

$$\eta(\mu) + W^\mu(x,a) = r(x,a)^2 + \sum_{x'} P(x'|x,a)U^\mu(x'). \quad (4)$$

Lemma 1 *Under Assumptions (A1) and (A2), we have*

$$\nabla\rho(\theta) = \sum_{x,a}\pi(x,a;\theta)\nabla\log\mu(a|x;\theta)Q(x,a;\theta), \quad (5)$$

$$\nabla\eta(\theta) = \sum_{x,a}\pi(x,a;\theta)\nabla\log\mu(a|x;\theta)W(x,a;\theta). \quad (6)$$

Proof. See Appendix A in the supplementary material. ■

Note that Eq. 6 for calculating $\nabla\eta(\theta)$ has close resemblance to Eq. 5 for $\nabla\rho(\theta)$, and thus, similar to what we have for Eq. 5, any function $b : \mathcal{X} \to \mathbb{R}$ can be added or subtracted to $W(x,a;\theta)$ on the RHS of Eq. 6 without changing the result of the integral (see e.g., Bhatnagar et al. 2009). So, we can replace $W(x,a;\theta)$ with the square reward advantage function $B(x,a;\theta) = W(x,a;\theta) - U(x;\theta)$ on the RHS of (6) in the same manner as we can replace $Q(x,a;\theta)$ with the advantage function $A(x,a;\theta) = Q(x,a;\theta) - V(x;\theta)$ on the RHS (5) without changing the result of the integral. We define the temporal difference (TD) errors δ_t and ϵ_t for value and square value functions as

$$\delta_t = R(x_t,a_t) - \widehat{\rho}_{t+1} + \widehat{V}(x_{t+1}) - \widehat{V}(x_t),$$
$$\epsilon_t = R(x_t,a_t)^2 - \widehat{\eta}_{t+1} + \widehat{U}(x_{t+1}) - \widehat{U}(x_t).$$

If $\widehat{V}$, $\widehat{U}$, $\widehat{\rho}$, and $\widehat{\eta}$ are unbiased estimators of V^μ, U^μ, $\rho(\mu)$, and $\eta(\mu)$, respectively, then we can show that δ_t and ϵ_t are unbiased estimates of the advantage functions A^μ and B^μ.

Lemma 2 *For any given policy μ, we have*

$$\mathbb{E}[\,\delta_t|\,x_t,a_t,\mu] = A^\mu(x_t,a_t), \quad \mathbb{E}[\,\epsilon_t|\,x_t,a_t,\mu] = B^\mu(x_t,a_t).$$

Input: parameterized policy $\mu(\cdot|\cdot;\theta)$ and value function feature vectors $f(\cdot)$ and $g(\cdot)$
Initialization: policy parameters $\theta = \theta_0$; value function weight vectors $v = v_0$ and $u = u_0$; initial state $x_0 \sim P_0(x)$
for $t = 0, 1, 2, \ldots$ **do**
  Draw action $a_t \sim \mu(\cdot|x_t;\theta_t)$
  Observe next state $x_{t+1} \sim P(\cdot|x_t,a_t)$
  Observe reward $r(x_t,a_t)$
  (i) Average Updates:
  $$\widehat{\rho}_{t+1} = \big(1-d(t)\big)\widehat{\rho}_t + d(t)R(x_t,a_t)$$
  $$\widehat{\eta}_{t+1} = \big(1-d(t)\big)\widehat{\eta}_t + d(t)R(x_t,a_t)^2$$
  (ii) TD Errors:
  $$\delta_t = R(x_t,a_t) - \widehat{\rho}_{t+1} + v_t^\top f(x_{t+1}) - v_t^\top f(x_t)$$
  $$\epsilon_t = R(x_t,a_t)^2 - \widehat{\eta}_{t+1} + u_t^\top g(x_{t+1}) - u_t^\top g(x_t)$$
  (iii) Critic Update: see Eq. 7 in the text
  (iv) Actor Update: see Eqs. 8 and 9 in the text
end for
return policy and value function parameters θ, λ, v, u

Figure 1. Template of the Average Reward Actor-Critic Alg.

Proof. See Appendix A in the supplementary material. ■

From Lemma 2 we notice that $\delta_t\psi_t$ and $\epsilon_t\psi_t$ are unbiased estimates of $\nabla\rho(\mu)$ and $\nabla\eta(\mu)$, respectively, where $\psi_t = \psi(x_t,a_t) = \nabla\log\mu(a_t|x_t)$ is the *compatible* feature (see e.g., Sutton et al. 2000 and Peters et al. 2005).

We now present our risk-sensitive actor-critic algorithm for average reward MDPs. The algorithm is in the general form shown in Figure 1. Let $\widehat{V}(x) = v^\top f(x)$ and $\widehat{U}(x) = u^\top g(x)$ denote the parameterized approximation to the differential value functions (for reward and square reward) in state x. One can also denote the same as $\widehat{V} = \Phi_v v$ and $\widehat{U} = \Phi_u u$, where Φ_v and Φ_u are $n \times d_2$ and $n \times d_3$ dimensional matrices ($d_2, d_3 \ll n$) whose ith columns are $f_i = (f_i(x), x \in \mathcal{X})^\top$, $i = 1,\ldots,d_2$ and $g_i = (g_i(x), x \in \mathcal{X})^\top$, $i = 1,\ldots,d_3$. We make the following standard assumption as in Bhatnagar et al. (2009):

(A3) *The basis functions $\{f_i\}_{i=1}^{d_2}$ and $\{g_i\}_{i=1}^{d_3}$ are linearly independent. In particular, $d_2, d_3 \ll n$ and Φ_v and Φ_u are full rank. Moreover, for every $v \in \mathbb{R}^{d_2}$ and $u \in \mathbb{R}^{d_3}$, $\Phi_v v \neq e$ and $\Phi_u u \neq e$, where e is the n-dimensional vector with all entries equal to one.*

The algorithm is incremental and attempts to find a (local) saddle point of $L(\theta,\lambda)$ by performing the following update procedure at each time step t:[1]

(i) Average Updates: We first update the average reward $\widehat{\rho}_t$ and the average square reward $\widehat{\eta}_t$ as shown in Algorithm 1.
(ii) TD Errors: We then calculate the TD-errors δ_t and ϵ_t as shown in Algorithm 1.

[1] In case of optimizing the Sharpe ratio (SR), it attempts to find a local maximum of $S(\theta)$.

(iii) Critic Update: The critic uses TD and updates the differential value v_t and square value u_t function parameters

$$v_{t+1} = v_t + c(t)\delta_t f(x_t), \qquad u_{t+1} = u_t + c(t)\epsilon_t g(x_t). \quad (7)$$

(iv) Actor Update: The actor updates the policy parameters θ (Eq. 8) and the Lagrange multiplier λ as follows:[2]

$$\theta_{t+1} = \Gamma\Big(\theta_t - b(t)\big(-\delta_t\psi_t + \lambda_t(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big), \quad (8)$$

$$\lambda_{t+1} = \Gamma_\lambda\Big(\lambda_t + a(t)(\widehat{\eta}_{t+1} - \widehat{\rho}_{t+1}^2 - \alpha)\Big). \quad (9)$$

Γ is an operator that projects a vector $\theta \in \mathbb{R}^{d_1}$ (d_1 is the dimension of the policy parameters) to the closest point in a compact and convex set $C \subset \mathbb{R}^{d_1}$, and Γ_λ is a projection operator to $[0, \lambda_{\max}]$. These projection operators are necessary for the convergence proof of the algorithm. Finally we have the assumption (A4) for the four step-size schedules.

(A4) *The step size schedules $\{d(t)\}$, $\{c(t)\}$, $\{b(t)\}$, and $\{a(t)\}$ satisfy (k is some positive constant)*

$$\sum_t a(t) = \sum_t b(t) = \sum_t c(t) = \infty, \quad (10)$$

$$\sum_t a(t)^2,\ \sum_t b(t)^2,\ \sum_t c(t)^2 < \infty, \quad (11)$$

$$a(t) = o\big(b(t)\big),\ b(t) = o\big(c(t)\big),\ d(t) = kc(t). \quad (12)$$

Eqs. 10 and 11 are standard step-size conditions in stochastic approximation algorithms, and Eq. 12 indicates that the average and critic updates are on the (same) fastest time-scale $\{d(t)\}$ and $\{c(t)\}$, the policy parameters update is on the intermediate time-scale $\{b(t)\}$, and the Lagrange multiplier is on the slowest time-scale $\{a(t)\}$. This means that the proposed algorithm is a three time-scale stochastic approximation algorithm. Note that we only need a two time-scale algorithm to optimize SR (no λ-recursion).

Although our estimates of $\rho(\theta)$ and $\eta(\theta)$ are unbiased, we use biased estimates for V^θ and U^θ, and thus, our gradient estimates $\nabla\rho(\theta)$ and $\nabla\eta(\theta)$ are biased. The following lemma shows the bias in our estimate of $\nabla L(\theta, \lambda)$.

Lemma 3 *The bias of our actor-critic algorithm in estimating $\nabla L(\theta, \lambda)$ for fixed θ and λ is*

$$\mathcal{B}(\theta,\lambda) = \sum_x d^\theta(x)\Big\{-\big(1 + 2\lambda\rho(\theta)\big)\big[\nabla\bar{V}^\theta(x) - \nabla {v^\theta}^\top f(x)\big] + \lambda\big[\nabla\bar{U}^\theta(x) - \nabla {u^\theta}^\top g(x)\big]\Big\},$$

where ${v^\theta}^\top f(\cdot)$ and ${u^\theta}^\top g(\cdot)$ are estimates of $V^\theta(\cdot)$ and $U^\theta(\cdot)$ upon convergence of the TD recursion, and

$$\bar{V}^\theta(x) = \sum_a \mu(a|x)\big[r(x,a) - \rho(\theta) + \sum_{x'} P(x'|x,a){v^\theta}^\top f(x')\big],$$

$$\bar{U}^\theta(x) = \sum_a \mu(a|x)\big[r(x,a)^2 - \eta(\theta) + \sum_{x'} P(x'|x,a){u^\theta}^\top g(x')\big].$$

[2] See Appendix D in the supplementary material for the actor's update formula (and other formulations) for SR.

Proof. See Appendix A in the supplementary material. ■

The convergence analysis of our algorithm follows the ordinary differential equation (ODE) approach (Borkar, 2008), and has three main steps. We report the detailed proof in Appendix A in the supplementary material. The first step is the convergence of ρ, η, V, and U, for any fixed policy θ and Lagrange multiplier λ. This corresponds to a TD(0) (with extension to η and U) proof. The policy and Lagrange multiplier are considered fixed because the critic's updates are on the faster time-scale than the actor's. The second step is to show the convergence of θ_t to an ε-neighborhood $\mathcal{Z}_\lambda^\varepsilon$ of the set of asymptotically stable equilibria $\mathcal{Z}_\lambda$ of ODE

$$\dot{\theta}_t = \check{\Gamma}\big(-\nabla L(\theta_t, \lambda)\big), \quad (13)$$

where for any bounded continuous function $\zeta(\cdot)$, the projection operator $\check{\Gamma}$ is defined as

$$\check{\Gamma}\big(\zeta(\theta_t)\big) = \lim_{\tau\to 0}\frac{\Gamma\big(\theta_t + \tau\zeta(\theta_t)\big) - \theta_t}{\tau}. \quad (14)$$

$\check{\Gamma}$ ensures that the evolution of θ via the ODE (13) stays within the compact and convex set $C \subset \mathbb{R}^{d_1}$. Again here it is assumed that λ is fixed because θ-recursion is on a faster time-scale than λ's. Finally the third step is the convergence of λ and showing that the whole algorithm converges to a local saddle point of $L(\theta, \lambda)$.

5. Discounted Reward Setting

In a discounted reward MDP, for a given policy μ, we define the (discounted) return of a state x (state-action pair (x, a)) as the sum of (discounted) rewards encountered by the agent when it starts at state x (state-action pair (x, a)) and then follows policy μ, i.e.,

$$D^\mu(x) = \sum_{t=0}^\infty \gamma^t R(x_t, a_t) \mid x_0 = x,\ \mu,$$

$$D^\mu(x,a) = \sum_{t=0}^\infty \gamma^t R(x_t, a_t) \mid x_0 = x,\ a_0 = a,\ \mu.$$

The expected value of the above two random variables are the value and action-value functions of policy μ, i.e., $V^\mu(x) = \mathbb{E}\big[D^\mu(x)\big]$ and $Q^\mu(x) = \mathbb{E}\big[D^\mu(x,a)\big]$. The goal in the standard (discounted) reward formulation is to find an optimal policy $\mu^* = \arg\max_\mu V^\mu(x^0)$, where x^0 is the initial state of the system.[3]

Measuring the *variability* in the stream of rewards is more difficult in discounted than average reward MDPs. The most common measure is the *variance of the return*

$$\Lambda^\mu(x) = \mathbb{E}\big[D^\mu(x)^2\big] - V^\mu(x)^2 = U^\mu(x) - V^\mu(x)^2, \quad (15)$$

first introduced by Sobel (1982). Note that $U^\mu(x)$ in Eq. 15 is the *square reward value function* of state x under policy μ. Another popular variability measure is the *discounted normalized variance* (Filar et al., 1989)

[3] This can be easily extended to the case that the system has more than one initial state $\mu^* = \arg\max_\mu \int dx\ V^\mu(x)P_0(x)$.

$$\Lambda(\mu) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t\big(R_t - \rho_\gamma(\mu)\big)^2\right], \tag{16}$$

where $\rho_\gamma(\mu) = \sum_{x,a} d_\gamma^\mu(x|x^0)\mu(x|a)r(x,a)$ and $d_\gamma^\mu(x|x^0)$ is the discounted visiting distribution of state x under policy μ, defined as

$$d_\gamma^\mu(x|x^0) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \Pr(x_t = x|x_0 = x^0;\mu).$$

We can also define discounted visiting distribution of state-action pair (x,a) under policy μ as $\pi_\gamma^\mu(x,a|x^0) = d_\gamma^\mu(x|x^0)\mu(a|x)$.

The variability measure of Eq. 16 has close resemblance to the average reward variability measure of Eq. 2, and thus, any (discounted) risk measure based on Eq. 16 can be optimized similar to the corresponding average reward risk measure based on Eq. 2. Therefore, we only consider the variability measure of Eq. 15 for discounted MDPs in this paper. Although Λ^μ of Eq. 15 satisfies a Bellman equation, unfortunately, it lacks the monotonicity property of dynamic programming (DP), and thus, it is not clear how the related risk measures can be optimized by standard DP algorithms (Sobel, 1982). This is why policy gradient and actor-critic algorithms are good candidates to deal with this risk-sensitive performance measure. We consider the following risk-sensitive measure for discounted MDPs:

$$\max_\theta V^\theta(x^0) \qquad \text{subject to} \qquad \Lambda^\theta(x^0) \le \alpha. \tag{17}$$

The goal would be to find a policy μ^* (or θ^*) that maximizes (17). Similar to the average reward case, the proposed algorithms can be easily extended to other similar variance-based risk measures.

6. Discounted Reward Algorithms

In this section, we propose actor-critic algorithms for optimizing the risk-sensitive measure (17) when the measure of variability is defined by Eq. 15. We use the same Lagrangian relaxation procedure as in Section 4 to convert the constrained optimization problem (17) to an unconstrained one. In order to optimize (17), we should update the policy parameters in the direction of the gradients of $L(\theta,\lambda)$ that is now defined as

$$\nabla L(\theta,\lambda) = -\nabla V^\theta(x^0) + \lambda\nabla\Lambda^\theta(x^0). \tag{18}$$

Since $\nabla\Lambda(x^0) = \nabla U(x^0) - 2V(x^0)\nabla V(x^0)$, in order to compute $\nabla\Lambda(x^0)$ it would be enough to calculate $\nabla U(x^0)$. From the Bellman equation of $\Lambda^\mu(x)$, proposed by Sobel (1982), it is straightforward to derive the following Bellman equations for $U^\mu(x)$ and $W^\mu(x,a) = \mathbb{E}\big[D^\mu(x,a)^2\big]$:

$$\begin{aligned}
U^\mu(x) &= \sum_a \mu(a|x)r(x,a)^2 + \gamma^2\sum_{a,x'}\mu(a|x)P(x'|x,a)U^\mu(x') \\
&\quad + 2\gamma\sum_{a,x'}\mu(a|x)P(x'|x,a)r(x,a)V^\mu(x'), \\
W^\mu(x,a) &= r(x,a)^2 + \gamma^2\sum_{x'}P(x'|x,a)U^\mu(x') \\
&\quad + 2\gamma r(x,a)\sum_{x'}P(x'|x,a)V^\mu(x').
\end{aligned} \tag{19}$$

Using these definitions and notations we are now ready to derive expressions for the gradient of $V(x^0)$ and $U(x^0)$ that are the main ingredient in calculating $\nabla L(\theta,\lambda)$.

Lemma 4 *Under Assumptions (A1), we have*

$$(1-\gamma)\nabla V^\theta(x^0) = \sum_{x,a}\pi_\gamma^\theta(x,a|x^0)\nabla\log\mu(a|x;\theta)Q^\theta(x,a),$$

$$\begin{aligned}
(1-\gamma^2)\nabla U^\theta(x^0) &= \sum_{x,a}\widetilde{\pi}_\gamma^\theta(x,a|x^0)\nabla\log\mu(a|x;\theta)W^\theta(x,a) \\
&\quad + 2\gamma\sum_{x,a,x'}\widetilde{\pi}_\gamma^\theta(x,a|x^0)P(x'|x,a)r(x,a)\nabla V^\theta(x'),
\end{aligned}$$

where $\widetilde{\pi}_\gamma^\theta(x,a|x^0) = \widetilde{d}_\gamma^\theta(x|x^0)\mu(a|x)$ *and*

$$\widetilde{d}_\gamma^\theta(x|x^0) = (1-\gamma^2)\sum_{t=0}^{\infty}\gamma^{2t}\Pr(x_t = x|x_0 = x^0).$$

Proof. See Appendix B in the supplementary material. ■

Unlike the average reward formulation, it is challenging to devise an efficient method to estimate $\nabla_\theta L(\theta,\lambda)$ using the gradient formulas of Lemma 4. This is mainly because **1)** two different sampling distributions are used for ∇V^θ and ∇U^θ, and **2)** $\nabla V^\theta(x')$ appears in the second sum of ∇U^θ equation. This is why we use simultaneous perturbation methods for estimating the gradients in this paper and leave directly dealing with the gradient equations for future work.

6.1. Simultaneous Perturbation Algorithms

In this section, we present actor-critic algorithms for optimizing the risk-sensitive measure (17) that are based on two simultaneous perturbation methods: *simultaneous perturbation stochastic approximation* (SPSA) and *smoothed functional* (SF) (Bhatnagar et al., 2013). The proposed algorithms use simultaneous perturbation methods to estimate the gradient of the risk-sensitive objective function (18). The idea is to estimate the gradients $\nabla V^{\theta_t}(x^0)$ and $\nabla U^{\theta_t}(x^0)$ using two simulated trajectories of the system corresponding to policies θ_t and $\theta_t^+ = \theta_t + \beta\Delta_t$, where $\beta > 0$ is a positive constant and Δ_t is a perturbation random variable.The choice of the perturbation random variables, Δ_t's, is specific to the algorithm: Rademacher random variables for SPSA and Gaussian for SF.

SPSA-based gradient estimates were first proposed in Spall (1992) and have been widely studied and found to be

highly efficient in various settings, especially those involving high-dimensional parameters. The idea is to estimate the gradient of a function $H(\theta)$ as

$$\widehat{\partial_{\theta^{(i)}} H}(\theta) = \frac{H(\theta+\beta\Delta) - H(\theta-\beta\Delta)}{\beta\Delta^{(i)}}, i = 1, \ldots, d_1. \quad (20)$$

It can be shown that $\mathbb{E}\big[\widehat{\partial_{\theta^{(i)}} H}(\theta)\big] = \partial_{\theta^{(i)}} H(\theta) + O(\beta)$. The advantage of this estimator is that it perturbs all directions at the same time (the numerator is identical in all d_1 components). So, the number of function measurements needed for this estimator is always two, independent of the dimension d_1. However, unlike the SPSA estimates in Spall (1992) that use two-sided balanced estimates (simulations with parameters $\theta_t - \beta\Delta_t$ and $\theta_t + \beta\Delta_t$), our gradient estimates are one-sided (simulations with parameters θ_t and $\theta_t + \beta\Delta_t$) and resemble those in Chen et al. (1999). The use of one-sided estimates is primarily because for updates of the Lagrangian parameter λ_t we require a simulation with the running parameter θ_t. Using a balanced gradient estimate would therefore come at the cost of an additional simulation (the resulting procedure would then require three simulations) which we avoid in our method.

SF-based method estimates not the gradient of $\nabla H(\theta)$ itself, but rather the convolution of $\nabla H(\theta)$ with the probability density function of the Gaussian $\mathcal{N}(\mathbf{0}, \beta^2 \boldsymbol{I})$, i.e.,

$$C_\beta H(\theta) = \int \mathcal{G}_\beta(\theta - z) \nabla_z H(z) dz,$$

where $\mathcal{G}_\beta$ is a d_1-dimensional p.d.f. Using integration by parts, it is easy to see that

$$C_\beta H(\theta) = \int \nabla_z \mathcal{G}_\beta(z) H(\theta - z) dz.$$

Using the fact that $\nabla_z \mathcal{G}_\beta(z) = \frac{-z}{\beta^2}\mathcal{G}_\beta(z)$ and by substituting $z' = z/\beta$, we obtain

$$C_\beta H(\theta) = \frac{1}{\beta}\int -z' \mathcal{G}_1(z') H(\theta - \beta z') dz'.$$

As $\beta \to 0$, $C_\beta H(\theta)$ converges to $\nabla_\theta H(\theta)$. A one-sided SF estimate of $\nabla_\theta H(\theta)$ is given by

$$\nabla_\theta H(\theta) \approx \mathbb{E}\Big[\frac{\Delta}{\beta}\big(H(\theta+\beta\Delta) - H(\theta)\big)\Big], \quad (21)$$

where Δ is a d_1-vector of independent $\mathcal{N}(0,1)$ random variables (see Chapter 6 of Bhatnagar et al. 2013).

The proposed actor-critic algorithms based on SPSA and SF have the following form: at each time step t, the algorithms **1)** take action $a_t \sim \mu(\cdot|x_t;\theta_t)$ and observe the reward $r(x_t, a_t)$ and next state x_{t+1} in the first trajectory, **2)** take action $a_t^+ \sim \mu(\cdot|x_t^+;\theta_t^+)$ and observe the reward $r(x_t^+, a_t^+)$ and next state x_{t+1}^+ in the second trajectory, **3) Critic Update:** calculate the TD-errors δ_t, δ_t^+ for the value and ϵ_t, ϵ_t^+ for the square value functions (Eq. 23) and update the critic parameters v_t, v_t^+ for the value and u_t, u_t^+ for the square value functions (Eq. 22), **4) Actor Update:** estimate the gradients $\nabla V^\theta(x^0)$ and $\nabla U^\theta(x^0)$ using SPSA (the one-sided version of Eq. 20) or SF (Eq. 21) and update the policy parameters and the Lagrange multiplier in the direction of the estimated gradient (Eqs. 24-26). Figure 2 shows the overall flow of the algorithms.

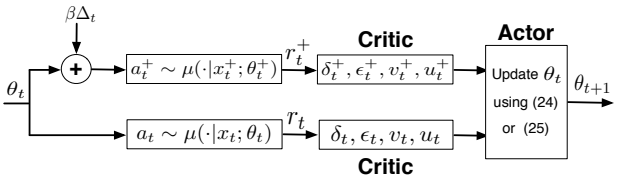


Figure 2. Overall flow of simultaneous perturbation algorithms.

Critic Update: We use the same linear representation for critic as in Section 4, i.e., $\widehat{V}(x) = v^\top f(x)$ and $\widehat{U}(x) = u^\top g(x)$, where $\{f_i\}_{i=1}^{d_2}$ and $\{g_i\}_{i=1}^{d_3}$ satisfy (A3). At each time step t, the critic updates the parameters of the value (v_t, v_t^+) and square value (u_t, u_t^+) functions for policies θ_t and $\theta_t^+ = \theta_t + \beta\Delta_t$ using TD as follows:

$$\begin{aligned} v_{t+1} &= v_t + c(t)\delta_t f(x_t), \quad v_{t+1}^+ = v_t^+ + c(t)\delta_t^+ f(x_t^+), \\ u_{t+1} &= u_t + c(t)\epsilon_t g(x_t), \quad u_{t+1}^+ = u_t^+ + c(t)\epsilon_t^+ g(x_t^+), \end{aligned} \quad (22)$$

where the TD-errors $\delta_t, \delta_t^+, \epsilon_t, \epsilon_t^+$ are computed as

$$\begin{aligned} \delta_t &= r(x_t, a_t) + \gamma v_t^\top f(x_{t+1}) - v_t^\top f(x_t), \\ \delta_t^+ &= r(x_t^+, a_t^+) + \gamma v_t^{+\top} f(x_{t+1}^+) - v_t^{+\top} f(x_t^+), \\ \epsilon_t &= r(x_t, a_t)^2 + 2\gamma r(x_t, a_t) v_t^\top f(x_{t+1}) + \gamma^2 u_t^\top g(x_{t+1}) \\ &\quad - u_t^\top g(x_t), \\ \epsilon_t^+ &= r(x_t^+, a_t^+)^2 + 2\gamma r(x_t^+, a_t^+) v_t^{+\top} f(x_{t+1}^+) + \gamma^2 u_t^{+\top} g(x_{t+1}^+) \\ &\quad - u_t^{+\top} g(x_t^+). \end{aligned} \quad (23)$$

This TD algorithm to learn the value and square value functions is a straightforward extension of the algorithm proposed by Tamar et al. (2013) to the discounted setting. Note that the TD-error ϵ for the square value function U comes directly from the Bellman equation for U in Eq. 19.

Actor Update: The actor updates the policy parameters θ and the Lagrange multiplier λ as follows:[4]

$$\theta_{t+1}^{(i)} = \Gamma_i\Big(\theta_t^{(i)} - b(t)\Big(-\big(1 + 2\lambda v_t^\top f(x^0)\big)\frac{(v_t^+ - v_t)^\top f(x^0)}{\beta\Delta_t^{(i)}} + \lambda\frac{(u_t^+ - u_t)^\top g(x^0)}{\beta\Delta_t^{(i)}}\Big)\Big), \; i = 1, \ldots, d_1 \quad \textbf{SPSA} \quad (24)$$

$$\theta_{t+1}^{(i)} = \Gamma_i\Big(\theta_t^{(i)} - b(t)\Big(\frac{-\Delta_t^{(i)}\big(1 + 2\lambda v_t^\top f(x^0)\big)}{\beta}(v_t^+ - v_t)^\top f(x^0) + \lambda\frac{\Delta_t^{(i)}}{\beta}(u_t^+ - u_t)^\top g(x^0)\Big)\Big), \; i = 1, \ldots, d_1 \quad \textbf{SF} \quad (25)$$

$$\lambda_{t+1} = \Gamma_\lambda\Big[\lambda_t + a(t)\Big(u_t^\top g(x^0) - \big(v_t^\top f(x^0)\big)^2 - \alpha\Big)\Big]. \quad (26)$$

[4]The policy parameters update for the Sharpe ratio (SR) objective function $S(\theta)$ may be derived similarly.

Note that 1) the λ-update is the same for both SPSA and SF methods, 2) $\Delta_t^{(i)}$'s are independent Rademacher and Gaussian random variables in SPSA and SF updates, respectively, 3) Γ and Γ_λ are projection operators similar to Eqs. 8 and 9, and 4) the step-size schedules $\{c(t)\}$, $\{b(t)\}$, and $\{a(t)\}$ satisfy (A4), meaning that the critic updates are on the fastest time-scale, the policy parameters update is on the intermediate time-scale, and the Lagrange multiplier update is on the slowest time-scale. Similar to the proposed average reward algorithm in Section 4, the SPSA and SF algorithms of this section are three (two in the case of Sharpe ratio) time-scale stochastic approximation algorithms.

Remark 1 *For the Sharpe ratio (SR) objective, the proposed SPSA and SF algorithms can be modified to work with only one simulated trajectory of the system. This is because in the SR case, we do not require to tune λ, and thus, the simulated trajectory corresponding to the nominal policy parameter θ is not necessary. In this implementation, the gradient is estimated as $\partial_{\theta^{(i)}} S(\theta) \approx S(\theta+\beta\Delta)/\beta\Delta^{(i)}$ for SPSA and as $\partial_{\theta^{(i)}} S(\theta) \approx (\Delta^{(i)}/\beta)S(\theta+\beta\Delta)$ for SF.*

Remark 2 *In the proposed algorithms, the critic uses a TD method to evaluate the policies. These algorithms can be implemented with a Monte-Carlo critic that at each time t computes a sample average of the total discounted rewards corresponding to the nominal θ_t and perturbed $\theta_t + \beta\Delta$ policy parameters. This implementation would be similar to that in Tamar et al. (2012), except here we use simultaneous perturbation methods to estimate the gradient.*

Remark 3 *Average reward analogues of our simultaneous perturbation algorithms can be developed. These algorithms would estimate the average reward ρ and the square reward η on the faster timescale and use these to estimate the gradient of the performance objective. However, a drawback with this approach, compared to the algorithm proposed in Section 4, is the necessity for having two simulated trajectories (instead of one) for each policy update.*

Sketch of the Convergence Analysis: The proof of convergence of the SPSA and SF algorithms to a (local) saddle point of the risk-sensitive objective function $\widehat{L}(\theta,\lambda) \stackrel{\triangle}{=} -\widehat{V}(\theta) + \lambda\big(\widehat{U}(\theta) - \widehat{V}^2(\theta)\big)$ contains the following three main steps. Note that since SPSA and SF use different methods to estimate the gradient, their proofs only differ in the second step, i.e., the convergence of the policy parameters θ. Due to space limitation, we only state the main theorems in the paper and report their proofs in Appendix C.

Step 1: (Critic's Convergence) The goal here is to show that the value and square value estimates of policies θ and $\theta^+ = \theta+\beta\Delta$ converge to the right values. Since the critic's update is on the fastest time-scale and the step-size schedules satisfy (A4), we can assume in this analysis that θ and λ are time invariant quantities.

Theorem 5 *Under (A1)-(A4), for any given policy parameter θ and Lagrange multiplier λ, the critic parameters $\{v_t\}$,$\{v_t^+\}$ and $\{u_t\}$,$\{u_t^+\}$ governed by recursions of Eq. 22, converge to $v_t \to \bar{v}, v_t^+ \to \bar{v}^+$ and $u_t \to \bar{u}, u_t^+ \to \bar{u}^+$, where $\bar{v}, \bar{v}^+$ and $\bar{u}, \bar{u}^+$ are the unique solutions to*

$$(\Phi_v^\top \boldsymbol{D}_\gamma^\theta \Phi_v)\bar{v} = \Phi_v^\top \boldsymbol{D}_\gamma^\theta T_v^\theta[\Phi_v\bar{v}],$$
$$(\Phi_v^\top \boldsymbol{D}_\gamma^{\theta^+} \Phi_v)\bar{v}^+ = \Phi_v^\top \boldsymbol{D}_\gamma^{\theta^+} T_v^{\theta^+}[\Phi_v\bar{v}^+],$$
$$(\Phi_v^\top \boldsymbol{D}_\gamma^\theta \Phi_u)\bar{u} = \Phi_v^\top \boldsymbol{D}_\gamma^\theta T_u^\theta[\Phi_u\bar{u}],$$
$$(\Phi_v^\top \boldsymbol{D}_\gamma^{\theta^+} \Phi_u)\bar{u}^+ = \Phi_v^\top \boldsymbol{D}_\gamma^{\theta^+} T_u^{\theta^+}[\Phi_u\bar{u}^+],$$

where $\boldsymbol{D}_\gamma^\theta$ and $\boldsymbol{D}_\gamma^{\theta^+}$ denote the diagonal matrices with entries $d_\gamma^\theta(x)$ and $d_\gamma^{\theta^+}(x)$ for all $x \in \mathcal{X}$, and $T_v^\theta, T_v^{\theta^+}$ and $T_u^\theta, T_u^{\theta^+}$ are the Bellman operators for value and square value functions of policies θ and θ^+, respectively. For any $y \in \mathbb{R}^{2n}$ such that $y = [y_v; y_u]$ and $y_v, y_u \in \mathbb{R}^n$, these operators are defined as $T_v^\theta y = \boldsymbol{r}^\theta + \gamma \boldsymbol{P}^\theta y_v$ and $T_u^\theta y = \boldsymbol{R}^\theta \boldsymbol{r}^\theta + 2\gamma \boldsymbol{R}^\theta \boldsymbol{P}^\theta y_v + \gamma^2 \boldsymbol{P}^\theta y_u$, where $\boldsymbol{r}^\theta$ and $\boldsymbol{P}^\theta$ are the reward vector and transition probability matrix of policy θ, and $\boldsymbol{R}^\theta = diag(\boldsymbol{r}^\theta)$.

Note that $[\Phi_v\bar{v}; \Phi_u\bar{u}]$ (the value and square value functions that the critic converges to) is the unique fixed point of the projected Bellman operator ΠT, where T contains both Bellman operators T_v and T_u for value and square value functions and Π contains both projections Π_v and Π_u into the linear spaces spanned by the columns of Φ_v and Φ_u (see Tamar et al. 2013 for more details).

Step 2: (Analysis of θ-recursion) The goal here is to show that the update of θ is in the direction of $\nabla\widehat{L}(\theta,\lambda)$ and it converges to a limiting set that depends on λ. Note that similar to Step 1, since θ-recursion is on a faster time-scale than λ's, we can assume that λ is constant in this analysis.

Consider the ordinary differential equation (ODE)

$$\dot{\theta}_t = \check{\Gamma}\big(-\nabla\widehat{L}(\theta_t,\lambda)\big) \tag{27}$$

where the projection operator $\check{\Gamma}$, defined by Eq. (14), ensures that the evolution of θ via the ODE (40) stays within set $C \subset \mathbb{R}^{d_1}$. Let $\mathcal{Z}_\lambda = \big\{\theta \in C : \check{\Gamma}\big(-\nabla\widehat{L}(\theta_t,\lambda)\big) = 0\big\}$ denote the set of asymptotically stable equilibrium points of the ODE (40) and $\mathcal{Z}_\lambda^\varepsilon$ denote the set of points in the ε-neighborhood of $\mathcal{Z}_\lambda$.

Theorem 6 *Under (A1)-(A4), for any given Lagrange multiplier λ and $\epsilon > 0$, there exists $\beta_0 > 0$ such that for all $\beta \in (0,\beta_0)$, $\theta_t \to \theta^* \in \mathcal{Z}_\lambda^\varepsilon$ almost surely.*

Step 3: (Analysis of λ-recursion and Convergence to a Local Saddle Point) The goal here is to first show that the λ-recursion converges and then to prove that the whole algorithm converges to a local saddle point of $\widehat{L}(\theta,\lambda)$.

We define the following ODE governing the evolution of λ

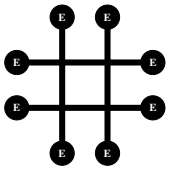


Figure 3. Road Network used for our Experiments.

$$\dot{\lambda}_t = \check{\Gamma}_\lambda\big(\widehat{U}(\theta_t) - \widehat{V}^2(\theta_t) - \alpha\big). \tag{28}$$

Theorem 7 *$\lambda_t \to \mathcal{F}$ almost surely as $t \to \infty$, where $\mathcal{F} \triangleq \{\lambda \mid \lambda \in [0, \lambda_{\max}], \check{\Gamma}_\lambda\big(\widehat{U}(\theta^\lambda) - \widehat{V}^2(\theta^\lambda) - \alpha\big) = 0,\ \theta^\lambda \in \mathcal{Z}_\lambda\}$.*

The last step is to establish that the algorithm converges to a (local) saddle point of $-\widehat{L}(\theta, \lambda)$, in other words, to a pair (θ^*, λ^*) that are a local minimum w.r.t. θ and a local maximum w.r.t. λ of $\widehat{L}$. From Theorem 7, $\lambda_t \to \lambda^*$ for some $\lambda^* \in [0, \lambda_{\max}]$ such that $\theta^{\lambda^*} \in Z_{\lambda^*}$ and $\check{\Gamma}_\lambda\big(\widehat{U}(\theta^{\lambda^*}) - \widehat{V}^2(\theta^{\lambda^*}) - \alpha\big) = 0$. We now invoke the envelope theorem of mathematical economics (Mas-Colell et al., 1995) to conclude that the ODE $\dot{\lambda}_t = \check{\Gamma}_\lambda\big(\Lambda(\theta_t) - \alpha\big)$ is equivalent to $\dot{\lambda}_t = \check{\Gamma}_\lambda\big(\nabla_\lambda \widehat{L}(\theta^{\lambda^*}, \lambda^*)\big)$. From the above, it is clear that (θ_t, λ_t) governed by Eqs. 24-26 converges to a local saddle point of $\widehat{L}(\cdot, \cdot)$.

7. Experimental Results

We evaluate our algorithms in the context of a traffic signal control application. The problem here is to adaptively choose the sign configuration that maximizes the traffic flow in the long-term. The objective in our formulation is to minimize the total number of vehicles in the system, which indirectly minimizes the delay experienced by the system. However, a traffic light control strategy that is not risk-sensitive may result in large variations in the delay experienced by road users. We demonstrate through our experiments that our risk-sensitive algorithms in comparison to their risk-neutral counterparts, result in significantly reduced variance in the mean sum of vehicles. Further, from the average waiting time experienced by the road users is only slightly higher in our algorithms.

We consider both infinite horizon average as well discounted settings for the traffic signal control MDP, formulated as in (Prashanth & Bhatnagar, 2011). We briefly recall their traffic control MDP formulation here: The state x_n is the vector of queue lengths and elapsed times and is given by $x_n = (q_1, \ldots, q_N, t_1, \ldots, t_N)$. Here q_i denotes the queue length on the signalled lane i of the network and t_i is the elapsed time since the signal turned to red on lane i. While the queue length factor is used to minimize the number of vehicles in the system, the elapsed time criterion ensures that a lane does suffer being red for a unfair amount of time. The actions a_n belong to the set of feasible sign configurations. The single-stage cost function $g(x_n)$ for the traffic control MDP is as follows:

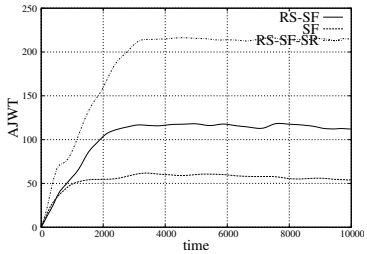


Figure 5. Performance Comparison of a SF-based alg. vs. RS-SF and RS-SF-SR algs. using average junction waiting time (AJWT).

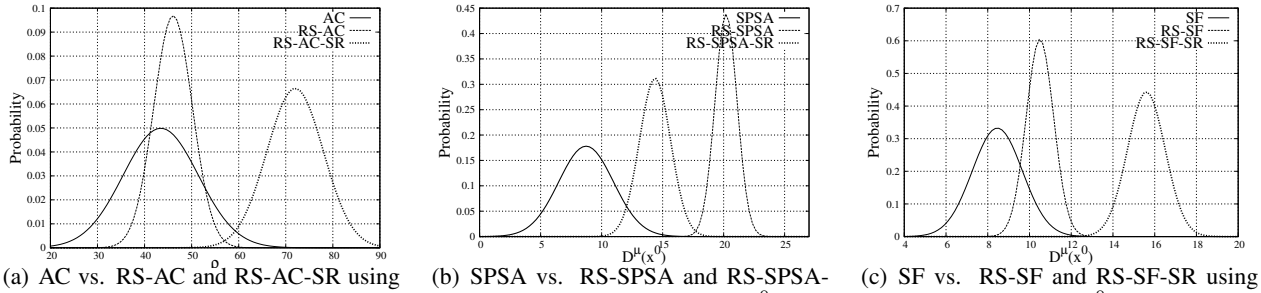
$$g(x_n, a_n) = r_1 * \Big(\sum_{i \in I_p} r_1 * q_i(n) + \sum_{i \notin I_p} s_2 * q_i(n)\Big) \tag{29}$$

$$+ s_1 * \Big(\sum_{i \in I_p} r_2 * t_i(n) + \sum_{i \notin I_p} s_2 * t_i(n)\Big), \tag{30}$$

where $r_i, s_i \geq 0$ and $r_i + s_i = 1, i = 1, 2$ and the set I_p is the set of prioritized lanes in the considered road network. While the weights r_1, s_1 are used to differentiate between the queue length and elapsed time factors, the weights r_2, s_2 help in prioritization of traffic. Given the above traffic control setting, we aim to minimize both the long run discounted as well average sum of the cost function $g(x_n)$. For the average setting, we implement a risk-neutral actor critic (AC) algorithm (similar to Alg. 1 of (Bhatnagar et al., 2009)) as well as two risk-sensitive algorithms - RS-AC and RS-AC-SR. While RS-AC corresponds to the algorithm described in Section 4 that attempts to solve (3), the RS-AC-SR algorithm consider the Sharpe ratio risk objective. On similar lines, we implement a plain actor critic algorithm (AC) for the discounted settings, as well as the risk-sensitive algorithms - RS-SPSA, RS- SF for solving (17) and the Sharpe variants RS-SPSA-SF and RS-SF, respectively. All the algorithms above use a parameterized Boltzmann policy (see Appendix E).

The road network used for our experiments is shown in Fig. 3. All our algorithms incorporate function approximation owing to the curse of dimensionality associated with larger road networks. For instance, assuming only 20 vehicles per lane of a 2x2-grid network, the cardinality of the state space is approximately of the order 10^{32} and the situation is aggravated as the size of the road network increases. The choice of features used in each of our algorithms is as described in Section V-B of (Prashanth & Bhatnagar, 2012). The detailed list of parameters and step-sizes chosen for our algorithms is given in Appendix E.

Fig. 4(a) shows the distribution of the average reward ρ for the algorithms in the average setting, while Figs. 4(b)–4(c) shows the distribution of the discounted cumulative reward $D^\mu(x^0)$ for the discounted setting. From these plots, we notice that the risk-sensitive algorithms that we propose result in a long-term (average or discounted) reward that is slightly higher than their risk-neutral variants. How-



(a) AC vs. RS-AC and RS-AC-SR using the distribution of ρ

(b) SPSA vs. RS-SPSA and RS-SPSA-SR using the distribution of $D^\mu(x^0)$

(c) SF vs. RS-SF and RS-SF-SR using the distribution of $D^\mu(x^0)$

Figure 4. Comparison of our risk-sensitive algorithms vs. their risk-neutral counterparts in both average as well as discounted settings.

ever, the empirical variance of the reward (both average as well as discounted) is observed to be significantly lesser with our proposed algorithms. Further, from a traffic signal control application standpoint, the performance of our risk-sensitive algorithms is very close to that of their risk-neutral counterparts. This is illustrated in the average junction waiting time performance plots in Fig. 5 for the SF algorithm and its risk sensitive variants (see Appendix E for similar results for the SPSA and average reward).

References

Bertsekas, D. *Nonlinear programming*. Athena Scientific, 1999.

Bhatnagar, S. An actor–critic algorithm with function approximation for discounted cost constrained Markov decision processes. *Systems & Control Letters*, 59(12):760–766, 2010.

Bhatnagar, S., Sutton, R., Ghavamzadeh, M., and Lee, M. Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482, 2009.

Bhatnagar, S., Prasad, H., and Prashanth, L. *Stochastic Recursive Algorithms for Optimization*, volume 434. Springer, 2013.

Borkar, V. A sensitivity formula for the risk-sensitive cost and the actor-critic algorithm. *Systems & Control Letters*, 44:339–346, 2001.

Borkar, V. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27:294–311, 2002.

Borkar, V. *Stochastic approximation: a dynamical systems viewpoint*. Cambridge University Press, 2008.

Chen, H., Duncan, T., and Pasik-Duncan, B. A Kiefer-Wolfowitz algorithm with randomized differences. *IEEE Transactions on Automatic Control*, 44(3):442–453, 1999.

Delage, E. and Mannor, S. Percentile optimization for Markov decision processes with parameter uncertainty. *Operations Research*, 58(1):203–213, 2010.

Filar, J., Kallenberg, L., and Lee, H. Variance-penalized Markov decision processes. *Mathematics of Operations Research*, 14 (1):147–161, 1989.

Filar, J., Krass, D., and Ross, K. Percentile performance criteria for limiting average Markov decision processes. *IEEE Transaction of Automatic Control*, 40(1):2–10, 1995.

Howard, R. and Matheson, J. Risk sensitive Markov decision processes. *Management Science*, 18(7):356–369, 1972.

Katkovnik, V. and Kulchitsky, Y. Convergence of a class of random search algorithms. *Automatic Remote Control*, 8:81–87, 1972.

Konda, V. and Tsitsiklis, J. Actor-Critic algorithms. In *Proceedings of Advances in Neural Information Processing Systems 12*, pp. 1008–1014, 2000.

Kushner, H. and Clark, D. *Stochastic approximation methods for constrained and unconstrained systems*. Springer-Verlag, 1978.

Mas-Colell, A., Whinston, M., and Green, J. *Microeconomic theory*. Oxford University Press, 1995.

Nilim, A. and Ghaoui, L. El. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

Peters, J., Vijayakumar, S., and Schaal, S. Natural actor-critic. In *Proceedings of the Sixteenth European Conference on Machine Learning*, pp. 280–291, 2005.

Prashanth, L.A. and Bhatnagar, S. Reinforcement Learning With Function Approximation for Traffic Signal Control. *IEEE Transactions on Intelligent Transportation Systems*, 12(2):412 –421, june 2011.

Prashanth, L.A. and Bhatnagar, S. Threshold Tuning Using Stochastic Optimization for Graded Signal Control. *IEEE Transactions on Vehicular Technology*, 61(9):3865 –3880, nov. 2012.

Puterman, M. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 1994.

Sharpe, W. Mutual fund performance. *Journal of Business*, 39(1): 119–138, 1966.

Sobel, M. The variance of discounted Markov decision processes. *Applied Probability*, pp. 794–802, 1982.

Spall, J. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Transactions on Automatic Control*, 37(3):332–341, 1992.

Sutton, R., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of Advances in Neural Information Processing Systems 12*, pp. 1057–1063, 2000.

Tamar, A., Castro, D. Di, and Mannor, S. Policy gradients with variance related risk criteria. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 387–396, 2012.

Tamar, A., Castro, D. Di, and Mannor, S. Policy evaluation with variance related risk criteria in markov decision processes. *arXiv preprint arXiv:1301.0104*, 2013.

Wiering, M., Vreeken, J., van Veenen, J., and Koopman, A. Simulation and optimization of traffic in a city. In *IEEE Intelligent Vehicles Symposium*, pp. 453–458, June 2004.

Xu, H. and Mannor, S. Distributionally robust Markov decision processes. *Mathematics of Operations Research*, 37(2):288–300, 2012.

Supplementary Material

A. Proofs of Section 4

Proof. **(Lemma 1)** Proof of $\nabla\rho(\theta)$ can be found in Sutton et al. (2000) and Konda & Tsitsiklis (2000). To prove $\nabla\eta(\theta)$, we start by the fact that from Eq. 4 we have $U(x) = \sum_a \mu(x|a)W(x,a)$. If we take the derivative w.r.t. θ from both sides of this equation, we obtain

$$\begin{aligned}
\nabla U(x) &= \sum_a \nabla\mu(x|a)W(x,a) + \sum_a \mu(x|a)\nabla W(x,a) \\
&= \sum_a \nabla\mu(x|a)W(x,a) + \sum_a \mu(x|a)\nabla\big(r(x,a)^2 - \eta + \sum_{x'} P(x'|x,a)U(x')\big) \\
&= \sum_a \nabla\mu(x|a)W(x,a) - \nabla\eta + \sum_{a,x'} \mu(a|x)P(x'|x,a)\nabla U(x'). \qquad (31)
\end{aligned}$$

The second equality is by replacing $W(x,a)$ from Eq. 4. Now if we take the weighted sum, weighted by $d(x)$, from both sides of Eq. 31, we have

$$\sum_x d(x)\nabla U(x) = \sum_{x,a} d(x)\nabla\mu(a|x)W(x,a) - \nabla\eta + \sum_{a,x'} d(x)\mu(a|x)P(x'|x,a)\nabla U(x'). \qquad (32)$$

The claim follows from the fact that the last sum on the RHS of Eq. 32 is equal to $\sum_x d(x)\nabla U(x)$. ■

Proof. **(Lemma 2)** The first statement $\mathbb{E}[\,\delta_t|\,x_t,a_t,\mu] = A^\mu(x_t,a_t)$ has been proved in Lemma 3 of Bhatnagar et al. (2009), so here we only prove the second statement $\mathbb{E}[\,\epsilon_t|\,x_t,a_t,\mu] = B^\mu(x_t,a_t)$. we may write

$$\begin{aligned}
\mathbb{E}[\,\epsilon_t|\,x_t,a_t,\mu] &= \mathbb{E}\big[R(x_t,a_t)^2 - \widehat{\eta}_{t+1} + \widehat{U}(x_{t+1}) - \widehat{U}(x_t) \mid x_t,a_t,\mu\big] \\
&= r(x_t,a_t)^2 - \eta(\mu) + \mathbb{E}\big[\widehat{U}(x_{t+1}) \mid x_t,a_t,\mu\big] - U^\mu(x_t) \\
&= r(x_t,a_t)^2 - \eta(\mu) + \mathbb{E}\Big[\mathbb{E}\big[\widehat{U}(x_{t+1}) \mid x_{t+1},\mu\big] \mid x_t,a_t\Big] - U^\mu(x_t) \\
&= r(x_t,a_t)^2 - \eta(\mu) + \mathbb{E}\big[\widehat{U}(x_{t+1}) \mid x_t,a_t\big] - U^\mu(x_t) \\
&= \underbrace{r(x_t,a_t)^2 - \eta(\mu) + \sum_{x_{t+1}\in\mathcal{X}} P(x_{t+1}|x_t,a_t)U^\mu(x_{t+1})}_{W^\mu(x,a)} - U^\mu(x_t) = B^\mu(x,a).
\end{aligned}$$

■

Proof. **(Lemma 3)** The bias in estimating $\nabla L(\theta,\lambda)$ consists of the bias in estimating $\nabla\rho(\theta)$ and $\nabla\eta(\theta)$. Lemma 4 in Bhatnagar et al. (2009) shows the bias in estimating $\nabla\rho(\theta)$ as

$$\mathbb{E}[\delta_t^\theta\psi_t|\theta] = \nabla\rho(\theta) + \sum_{x\in\mathcal{X}} d^\theta(x)\big[\nabla\bar{V}^\theta(x) - \nabla v^{\theta\top}f(x)\big],$$

where $\delta_t^\theta = R(x_t,a_t) - \widehat{\rho}_{t+1} + v^{\theta\top}f(x_{t+1}) - v^{\theta\top}f(x_t)$. Similarly we can prove that the bias in estimating $\nabla\eta(\theta)$ is

$$\mathbb{E}[\epsilon_t^\theta\psi_t|\theta] = \nabla\eta(\theta) + \sum_{x\in\mathcal{X}} d^\theta(x)\big[\nabla\bar{U}^\theta(x) - \nabla u^{\theta\top}g(x)\big],$$

where $\epsilon_t^\theta = R(x_t,a_t) - \widehat{\eta}_{t+1} + u^{\theta\top}g(x_{t+1}) - u^{\theta\top}g(x_t)$. The claim follows by putting these two results together and given the fact that $\nabla\Lambda(\theta) = \nabla\eta(\theta) - 2\rho(\theta)\nabla\rho(\theta)$ and $\nabla L(\theta,\lambda) = -\nabla\rho(\theta) + \lambda\nabla\Lambda(\theta)$.

Note that the following fact holds for the bias in estimating $\nabla\rho(\theta)$ and $\nabla\eta(\theta)$

$$\sum_x d^\theta(x)\big[\bar{V}^\theta(x) - v^{\theta\top} f(x)\big] = 0, \qquad \sum_x d^\theta(x)\big[\bar{U}^\theta(x) - u^{\theta\top} g(x)\big] = 0.$$

■

Convergence Analysis of Algorithm 1

Step 1: Critic's Convergence

Lemma 8 *For any given policy μ, $\{\widehat{\rho}_t\}$, $\{\widehat{\eta}_t\}$, $\{v_t\}$, and $\{u_t\}$ defined in the Algorithm 1 and by the critic recursion of Eq. 7, converge to $\rho(\mu)$, $\eta(\mu)$, v^μ, and u^μ with probability one, where v^μ and u^μ are the unique solution to*

$$\Phi_v^\top \boldsymbol{D}^\mu \Phi_v v^\mu = \Phi_v^\top \boldsymbol{D}^\mu T_v^\mu(\Phi_v v^\mu), \qquad \Phi_u^\top \boldsymbol{D}^\mu \Phi_u u^\mu = \Phi_u^\top \boldsymbol{D}^\mu T_u^\mu(\Phi_u u^\mu), \tag{33}$$

respectively. In Eq. 33, $\boldsymbol{D}^\mu$ denote the diagonal matrix with entries $d^\mu(x)$ for all $x \in \mathcal{X}$, and T_v^μ and T_u^μ are the Bellman operators for differential value and square value functions of policy μ, defined by Eq. 34, respectively.

$$T_v^\mu J = \boldsymbol{r}^\mu - \rho(\mu)\boldsymbol{e} + \boldsymbol{P}^\mu J, \qquad T_u^\mu J = \boldsymbol{R}^\mu \boldsymbol{r}^\mu - \eta(\mu)\boldsymbol{e} + \boldsymbol{P}^\mu J, \tag{34}$$

where $\boldsymbol{r}^\mu$ and $\boldsymbol{P}^\mu$ are reward vector and transition probability matrix of policy μ, $\boldsymbol{R}^\mu = diag(\boldsymbol{r}^\mu)$, and $\boldsymbol{e}$ is a vector of size n with elements all equal to one.

Proof. **(Lemma 8)** The proof follows the same steps as Lemma 5 in Bhatnagar et al. (2009). ■

Step 2: Actor's Convergence

Lemma 9 *Under Assumptions (A1)-(A4), given $\varepsilon > 0,\ \exists\delta > 0$ such that for $\theta_t,\ t \geq 0$ obtained using Algorithm 1, if $\sup_\theta \|\mathcal{B}(\theta,\lambda)\| < \delta$ then $\theta_t \to \mathcal{Z}_\lambda^\varepsilon$ as $t \to \infty$ with probability one.*

Proof. **(Lemma 9)** First note that the bias of Algorithm 1 in estimating $\nabla L(\theta,\lambda)$ is (see Lemma 3)

$$\mathcal{B}(\theta,\lambda) = \sum_x d^\theta(x)\Big\{ -\big(1 + 2\lambda\rho(\theta)\big)\big[\nabla\bar{V}^\theta(x) - \nabla v^{\theta\top} f(x)\big] + \lambda\big[\nabla\bar{U}^\theta(x) - \nabla u^{\theta\top} g(x)\big]\Big\}.$$

Also note that $\mathcal{Z}_\lambda = \big\{\theta \in C : \check{\Gamma}\big(-\nabla L(\theta,\lambda)\big) = 0\big\}$ denote the set of asymptotically stable equilibrium points of the ODE (40) and $\mathcal{Z}_\lambda^\varepsilon = \big\{\theta \in C : ||\theta - \theta_0|| < \varepsilon, \theta_0 \in \mathcal{Z}_\lambda\big\}$ denote the set of points in the ε-neighborhood of $\mathcal{Z}_\lambda$.

Let $\mathcal{F}(t) = \sigma(\theta_r, r \leq t)$ denote the sequence of σ-fields generated by $\theta_r,\ r \geq 0$. We have

$$\begin{aligned}
\theta_{t+1} &= \Gamma\Big(\theta_t - b(t)\big(-\delta_t\psi_t + \lambda(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)\big)\Big) \\
&= \Gamma\big(\theta_t + b(t)(1 + 2\lambda\widehat{\rho}_{t+1})\delta_t\psi_t - b(t)\lambda\epsilon_t\psi_t\big) \\
&= \Gamma\Bigg(\theta_t - b(t)\Big[1 + 2\lambda\Big(\big(\widehat{\rho}_{t+1} - \rho(\theta_t)\big) + \rho(\theta_t)\Big)\Big]\mathbb{E}\big[\delta^{\theta_t}\psi_t|\mathcal{F}(t)\big] \\
&\qquad - b(t)\Big[1 + 2\lambda\Big(\big(\widehat{\rho}_{t+1} - \rho(\theta_t)\big) + \rho(\theta_t)\Big)\Big]\Big(\delta_t\psi_t - \mathbb{E}\big[\delta_t\psi_t|\mathcal{F}(t)\big]\Big) \\
&\qquad - b(t)\Big[1 + 2\lambda\Big(\big(\widehat{\rho}_{t+1} - \rho(\theta_t)\big) + \rho(\theta_t)\Big)\Big]\mathbb{E}\big[(\delta_t - \delta^{\theta_t})\psi_t|\mathcal{F}(t)\big] \\
&\qquad + b(t)\lambda\mathbb{E}\big[\epsilon^{\theta_t}\psi_t|\mathcal{F}(t)\big] + b(t)\lambda\Big(\epsilon_t\psi_t - \mathbb{E}\big[\epsilon_t\psi_t|\mathcal{F}(t)\big]\Big) + b(t)\lambda\mathbb{E}\big[(\epsilon_t - \epsilon^{\theta_t})\psi_t|\mathcal{F}(t)\big]\Bigg).
\end{aligned}$$

By setting $\xi_t = \widehat{\rho}_{t+1} - \rho(\theta_t)$, we may write the above equation as

$$\begin{aligned}\theta_{t+1} = \Gamma\Big(\theta_t - b(t)\big[1+2\lambda\big(\xi_t+\rho(\theta_t)\big)\big]\mathbb{E}\big[\delta^{\theta_t}\psi_t|\mathcal{F}(t)\big] \\ - b(t)\big[1+2\lambda\big(\xi_t+\rho(\theta_t)\big)\big]\underbrace{\Big(\delta_t\psi_t - \mathbb{E}\big[\delta_t\psi_t|\mathcal{F}(t)\big]\Big)}_{*} \\ - b(t)\big[1+2\lambda\big(\xi_t+\rho(\theta_t)\big)\big]\underbrace{\mathbb{E}\big[(\delta_t-\delta^{\theta_t})\psi_t|\mathcal{F}(t)\big]}_{+} \\ + b(t)\lambda\mathbb{E}\big[\epsilon^{\theta_t}\psi_t|\mathcal{F}(t)\big] + b(t)\lambda\underbrace{\Big(\epsilon_t\psi_t - \mathbb{E}\big[\epsilon_t\psi_t|\mathcal{F}(t)\big]\Big)}_{*} + b(t)\lambda\underbrace{\mathbb{E}\big[(\epsilon_t-\epsilon^{\theta_t})\psi_t|\mathcal{F}(t)\big]}_{+}\Big). \end{aligned} \tag{35}$$

Since Algorithm 1 uses an unbiased estimator for ρ, we have $\widehat{\rho}_{t+1} \to \rho(\theta_t)$, and thus, $\xi_t \to 0$. The terms $(+)$ asymptotically vanish in lieu of Lemma 8 (Critic convergence). Finally the terms $(*)$ can be seen to vanish using standard martingale arguments (cf. Theorem 2 of Bhatnagar et al. 2009). Thus, Eq. 35 can be seen to be equivalent in an asymptotic sense to

$$\theta_{t+1} = \Gamma\Big(\theta_t - b(t)\big[1+2\lambda\rho(\theta_t)\big]\mathbb{E}\big[\delta^{\theta_t}\psi_t|\mathcal{F}(t)\big] + b(t)\lambda\mathbb{E}\big[\epsilon^{\theta_t}\psi_t|\mathcal{F}(t)\big]\Big). \tag{36}$$

From Lemma 3 and the foregoing, Eq. 8 asymptotically tracks the stable fixed points of the ODE

$$\dot{\theta}_t = \check{\Gamma}\Big(-\nabla L(\theta_t,\lambda) - \mathcal{B}(\theta_t,\lambda)\Big). \tag{37}$$

So, if the bias $\sup_\theta \|\mathcal{B}(\theta,\lambda)\| \to 0$, the trajectories (37) converge to those of (40) uniformly on compacts for the same initial condition and the claim follows. ■

Step 3: λ Convergence and Overall Convergence of the Algorithm

The goal here is to first show that the λ-recursion converges and then to prove that the whole algorithm converges to a local saddle point of $L(\theta,\lambda)$. We define the following ODE governing the evolution of λ

$$\dot{\lambda}_t = \check{\Gamma}_\lambda\big(\Lambda(\theta_t) - \alpha\big). \tag{38}$$

Theorem 10 *$\lambda_t \to \mathcal{F}$ almost surely as $t \to \infty$, where $\mathcal{F} \stackrel{\triangle}{=} \{\lambda \mid \lambda \in [0,\lambda_{\max}], \check{\Gamma}_\lambda\big(\Lambda(\theta^\lambda) - \alpha\big) = 0,\ \theta^\lambda \in \mathcal{Z}_\lambda\}$.*

The last step is to establish that the algorithm converges to a (local) saddle point of $-\widehat{L}(\theta,\lambda)$, in other words, to a pair (θ^*,λ^*) that are a local minimum w.r.t. θ and a local maximum w.r.t. λ of $L(\cdot,\cdot)$. From Theorem 10, $\lambda_t \to \lambda^*$ for some $\lambda^* \in [0,\lambda_{\max}]$ such that $\theta^{\lambda^*} \in Z_{\lambda^*}$ and $\check{\Gamma}_\lambda\big(\Lambda(\theta^{\lambda^*}) - \alpha\big) = 0$. We now invoke the envelope theorem of mathematical economics (Mas-Colell et al., 1995) to conclude that the ODE (38) is equivalent to $\dot{\lambda}_t = \check{\Gamma}_\lambda\big(\nabla_\lambda L(\theta^{\lambda^*},\lambda^*)\big)$. From the above, it is clear that (θ_t,λ_t) governed by Eqs. 8 and 9 converges to a local saddle point of $L(\cdot,\cdot)$.

B. Proofs of Section 6

Proof. **(Lemma 4)** The proof of $\nabla V^\theta(x^0)$ can be found in the literature (e.g., Peters et al. 2005). To prove $\nabla U^\theta(x^0)$, we start by the fact that from Eq. 19 we have $U(x) = \sum_a \mu(x|a)W(x,a)$. If we take the derivative w.r.t. θ from both sides of this equation, we obtain

$$\begin{aligned}
\nabla U(x^0) &= \sum_a \nabla\mu(x^0|a)W(x^0,a) + \sum_a \mu(a|x^0)\nabla W(x^0,a) \\
&= \sum_a \nabla\mu(a|x^0)W(x^0,a) + \sum_a \mu(a|x^0)\nabla\Big[r(x^0,a)^2 + \gamma^2 \sum_{x'} P(x'|x^0,a)U(x') + 2\gamma r(x^0,a)\sum_{x'} P(x'|x^0,a)V(x')\Big] \\
&= \underbrace{\sum_a \nabla\mu(x^0|a)W(x^0,a) + 2\gamma \sum_{a,x'} \mu(a|x^0)r(x^0,a)P(x'|x^0,a)\nabla V(x')}_{h(x^0)} + \gamma^2 \sum_{a,x'} \mu(a|x^0)P(x'|x^0,a)\nabla U(x') \\
&= h(x^0) + \gamma^2 \sum_{a,x'} \mu(a|x^0)P(x'|x^0,a)\nabla U(x') \qquad (39) \\
&= h(x^0) + \gamma^2 \sum_{a,x'} \mu(a|x^0)P(x'|x^0,a)\nabla\Big[h(x') + \gamma^2 \sum_{a',x''} \mu(a'|x')P(x''|x',a')\nabla U(x'')\Big]
\end{aligned}$$

By unrolling the last equation using the definition of $\nabla U(x)$ from Eq. 39, we obtain

$$\begin{aligned}
\nabla U(x^0) &= \sum_{t=0}^{\infty} \gamma^{2t} \sum_x \Pr(x_t = x|x_0 = x^0)h(x) = \frac{1}{1-\gamma^2}\sum_x \widetilde{d}(x|x^0)h(x) \\
&= \frac{1}{1-\gamma^2}\Big[\sum_{x,a} \widetilde{d}(x|x^0)\mu(a|x)\nabla\log\mu(a|x)W(x,a) + 2\gamma \sum_{x,a,x'} \widetilde{d}(x|x^0)\mu(a|x)r(x,a)P(x'|x,a)\nabla V(x')\Big] \\
&= \frac{1}{1-\gamma^2}\Big[\sum_{x,a} \widetilde{\pi}(x,a|x^0)\nabla\log\mu(a|x)W(x,a) + 2\gamma \sum_{x,a,x'} \widetilde{\pi}(x,a|x^0)r(x,a)P(x'|x,a)\nabla V(x')\Big].
\end{aligned}$$

■

C. Proofs of Section 6.1

Our algorithms use multi-timescale stochastic approximation and we use the ODE approach (see Chapter 6 of Borkar 2008) to analyze them. We first provide a sketch of the convergence proof of the SPSA algorithm. Later, we describe the necessary modifications for the analysis of the SF algorithm.

Step 1: (Critic's Convergence)

Proof. **(Theorem 5)** The proof of this theorem follows similar steps as in the proof of Theorem 10 of Tamar et al. (2013). For our analysis, we need to extend the proof of Tamar et al. (2013) to discounted MDPs and to the case that the reward is a function of both states and actions (and not just states), which is straightforward. ■

Step 2: (Analysis of θ-recursion)

We show that the update of θ is equivalent to gradient descent for the function $\widehat{L}(\theta,\lambda) \stackrel{\triangle}{=} -\widehat{V}(\theta) + \lambda\big(\widehat{U}(\theta) - \widehat{V}^2(\theta)\big)$ and converges to a limiting set that depends on λ. Consider the ODE

$$\dot{\theta}(t) = \check{\Gamma}\left(-\nabla_\theta \widehat{L}(\theta_t,\lambda)\right), \tag{40}$$

where $\check{\Gamma}$ is defined as follows: For any bounded continuous function $\zeta(\cdot)$,

$$\check{\Gamma}\big(\zeta(\theta_t)\big) = \lim_{\tau\to 0} \frac{\Gamma\big(\theta_t + \tau\zeta(\theta_t)\big) - \theta_t}{\tau}. \tag{41}$$

The projection operator $\check{\Gamma}(\cdot)$ ensures that the evolution of θ via the ODE (40) stays within the bounded set $C \in \mathbb{R}^{d_1}$. Due to timescale separation, the value of λ (updated on a slower timescale) is assumed to be constant for the analysis of the θ-update.

Let $\mathcal{Z}_\lambda = \big\{\theta \in C : \check{\Gamma}\big(-\nabla\widehat{L}(\theta_t,\lambda)\big) = 0\big\}$ denote the set of asymptotically stable equilibrium points of the ODE (40) and $\mathcal{Z}_\lambda^\varepsilon = \big\{\theta \in C : ||\theta - \theta_0|| < \varepsilon, \theta_0 \in \mathcal{Z}_\lambda\big\}$ denote the set of points in the ε-neighborhood of $\mathcal{Z}_\lambda$.

Proof. **(Theorem 6 for SPSA)** Since the TD critic converges on the faster timescale, the θ-update in (24) can be rewritten using the converged TD-parameters $(\bar{v},\bar{u})$ and $(\bar{v}^+,\bar{u}^+)$ as follows:

$$\theta_{t+1}^{(i)} = \Gamma_i\Bigg(\theta_t^{(i)} - b(t)\Big(-\big(1 + 2\lambda\bar{v}^\top f(x^0)\big)\frac{(\bar{v}^+ - \bar{v})^\top f(x^0)}{\beta\Delta_t^{(i)}} + \lambda\frac{(\bar{u}^+ - \bar{u})^\top g(x^0)}{\beta\Delta_t^{(i)}} + \xi_{1,t}\Big)\Bigg),$$

where $\xi_{1,t} \to 0$ (convergence of TD in the critic and as a result convergence of the critic's parameters to $\bar{v},\bar{u},\bar{v}^+,\bar{u}^+$) in lieu of Theorem 5.

Next, we establish that $\mathbb{E}\left[\frac{(\bar{v}^+ - \bar{v})^\top f(x^0)}{\beta\Delta^{(i)}} \mid \theta,\lambda\right]$ is a biased estimator of $\nabla_\theta\widehat{V}(\theta)$, where the bias vanishes asymptotically.

$$\begin{aligned}\mathbb{E}\left[\frac{(\bar{v}^+ - \bar{v})^\top f(x^0)}{\beta\Delta^{(i)}} \mid \theta,\lambda\right] =& \partial_{\theta^{(i)}}\bar{v}^\top f(x^0) + \mathbb{E}\left[\sum_{j\neq i}\frac{\Delta^{(j)}}{\Delta^{(i)}}\partial_{\theta^{(j)}}\bar{v}^\top f(x^0) \mid \theta,\lambda\right] + \xi_{2,t} f(x^0)\\ &\to_{\beta\to 0}\ \partial_{\theta^{(i)}}\bar{v}^\top f(x^0).\end{aligned}$$

The first equality above follows by expanding using Taylor series, whereas the second step follows by using the fact that $\Delta_t^{(i)}$'s are independent symmetric Bernoulli random variables. On similar lines, it can be seen that

$$\mathbb{E}\left[\frac{(\bar{u}^+ - \bar{u})^\top f(x^0)}{\beta\Delta^{(i)}} \mid \theta,\lambda\right] \to_{\beta\to 0}\ \partial_{\theta^{(i)}}\bar{u}^\top g(x^0).$$

Thus, (24) can be seen to be a discretization of the ODE (40). Further, $\mathcal{Z}_\lambda$ is an asymptotically stable attractor for the ODE (40), with $\widehat{L}(\theta,\lambda)$ itself serving as a strict Lyapunov function. This can be inferred as follows:

$$\frac{d\widehat{L}(\theta,\lambda)}{dt} = \nabla_\theta \widehat{L}(\theta,\lambda)\dot{\theta} = \nabla_\theta \widehat{L}(\theta,\lambda)\check{\Gamma}\big(-\nabla_\theta \widehat{L}(\theta,\lambda)\big) < 0.$$

The claim now follows from Theorem 5.3.3, pp. 191-196 of Kushner & Clark (1978). ■

Proof. **(Theorem 6 for SF)** The analysis proceeds along similar lines as the SPSA algorithm, with steps 1 and 3 corresponding to the fastest (critic-recursion) and slowest (λ-recursion) timescales being identical. However, the analysis of θ-recursion differs, with the smoothed functional based gradient estimate (25) in place of the SPSA-based estimate (24).

As in the case of the SPSA algorithm, we rewrite the θ-update in (25) using the converged TD-parameters as follows

$$\theta_{t+1}^{(i)} = \Gamma_i\Big(\theta_t^{(i)} - b(t)\Big(\frac{-\Delta_t^{(i)}\big(1+2\lambda\bar{v}^\top f(x^0)\big)}{\beta}(\bar{v}^+ - \bar{v})^\top f(x^0) + \lambda\frac{\Delta_t^{(i)}}{\beta}(\bar{u}^+ - \bar{u})^\top g(x^0) + \xi_{1,t}\Big)\Big),$$

where $\xi_{1,t} \to 0$ (convergence of TD in the critic and as a result convergence of the critic's parameters to $\bar{v}, \bar{u}, \bar{v}^+, \bar{u}^+$) in lieu of Theorem 5. Next, we establish that $\mathbb{E}\left[\frac{\Delta^{(i)}}{\beta}(\bar{v}^+ - \bar{v})^\top f(x^0) \mid \theta, \lambda\right]$ is an asymptotically correct estimate of the gradient of $\widehat{V}(\theta)$ in the following:

$$\mathbb{E}\left[\frac{\Delta^{(i)}}{\beta}(\bar{v}^+ - \bar{v})^\top f(x^0) \mid \theta, \lambda\right] \to_{\beta\to 0} \partial_{\theta^{(i)}}\bar{v}^\top f(x^0).$$

The above follows in a similar manner as Proposition 10.2 of Bhatnagar et al. (2013). On similar lines, one can see that that

$$\mathbb{E}\left[\frac{\Delta^{(i)}}{\beta}(\bar{u}^+ - \bar{u})^\top f(x^0) \mid \theta, \lambda\right] \to_{\beta\to 0} \partial_{\theta^{(i)}}\bar{u}^\top g(x^0).$$

Thus, (25) can be seen to be a discretization of the ODE (40) and the rest of the analysis follows in a similar manner as in the SPSA proof. ■

Step 3: (Analysis of λ-recursion and Convergence to a Local Saddle Point)

Proof. **(Theorem 7)** The proof follows the same steps as in Theorem 3 in Bhatnagar (2010). ■

D. Sharpe Ratio $S(\theta)$ Optimization

Gradient of SR in Average Reward MDPs

$$\nabla S(\theta) = \frac{1}{\sqrt{\Lambda(\theta)}}\big(\nabla\rho(\theta) - \frac{\rho(\theta)}{2\Lambda(\theta)}\nabla\Lambda(\theta)\big).$$

Actor Update for SR Optimization in Average Reward MDPs

$$\theta_{t+1} = \Gamma\Big(\theta_t + \frac{b(t)}{\sqrt{\widehat{\eta}_{t+1} - \widehat{\rho}_{t+1}^2}}\big(\delta_t\psi_t - \frac{\widehat{\rho}_{t+1}(\epsilon_t\psi_t - 2\widehat{\rho}_{t+1}\delta_t\psi_t)}{2(\widehat{\eta}_{t+1} - \widehat{\rho}_{t+1}^2)}\big)\Big). \tag{42}$$

Gradient of SR in Discounted Reward MDPs

$$\nabla S(\theta) = \frac{1}{\sqrt{\Lambda^\theta(x^0)}}\big(\nabla V^\theta(x^0) - \frac{V^\theta(x^0)}{2\Lambda^\theta(x^0)}\nabla\Lambda^\theta(x^0)\big).$$

Actor Update for SR Optimization in Discounted Reward MDPs

SPSA

$$\theta_{t+1} = \Gamma\left(\theta_t + \frac{b(t)}{\sqrt{u_t^\top g(x^0) - \big(v_t^\top f(x^0)\big)^2}\beta\Delta_t^{(i)}}\left((v_t^+ - v_t)^\top f(x^0) - \frac{v_t^\top f(x^0)((u_t^+ - u_t)^\top g(x^0) - 2v_t^\top f(x^0)(v_t^+ - v_t)^\top f(x^0)}{2(u_t^\top g(x^0) - \big(v_t^\top f(x^0)\big)^2)}\right)\right). \tag{43}$$

SF

$$\theta_{t+1} = \Gamma\left(\theta_t + \frac{b(t)}{\sqrt{u_t^\top g(x^0) - \big(v_t^\top f(x^0)\big)^2}}\frac{\Delta_t^{(i)}}{\beta}\left((v_t^+ - v_t)^\top f(x^0) - \frac{v_t^\top f(x^0)((u_t^+ - u_t)^\top g(x^0) - 2v_t^\top f(x^0)(v_t^+ - v_t)^\top f(x^0)}{2(u_t^\top g(x^0) - \big(v_t^\top f(x^0)\big)^2)}\right)\right). \tag{44}$$

E. Additional Simulation Experiments

We implement the following algorithms using the Green Light District (GLD) simulator (Wiering et al., 2004):
Average Setting:

- **AC**: This is an actor critic algorithm that minimize the long run average sum of the single-stage cost function $g(x_t, a_t)$, without considering any risk criteria. This is similar to Algorithm 1 in Bhatnagar et al. (2009).
- **RS-AC**: This is the risk-sensitive actor critic algorithm that attempts to solve (3) and is described in Section 4.
- **RS-AC-SR**: This is another risk-sensitive actor critic algorithm that considers the Sharpe ratio based risk measure instead of the constrained problem (3) and updates the actor according to (42).

Discounted Setting:

- **SPSA**: This is an actor critic algorithm that minimize the long run discounted sum of the single-stage cost function $g(x_t, a_t)$, without considering any risk criteria. This is similar to Algorithm 1 in (Bhatnagar, 2010).
- **RS-SPSA**: This is the risk-sensitive actor critic algorithm that attempts to solve (3) and updates according to (24).
- **RS-SPSA-SR**: This is another risk-sensitive actor critic algorithm that considers the Sharpe ratio based risk measure instead of the constrained problem (17) and updates the actor according to (43).
- **SF**: This is similar to SPSA algorithm above, except that the gradient estimation scheme used here is based on the smoothed functional technique. The update of the policy parameter in this algorithm is given by

$$\theta_{t+1}^{(i)} = \Gamma_i \left(\theta_t^{(i)} + b(t) \Big(\frac{\Delta_t^{(i)}}{\beta} (v_t^+ - v_t)^\top f(x^0) \Big) \right).$$

- **RS-SF**: This is the risk-sensitive variant of the SF algorithm above and updates the actor according to (25).
- **RS-SPSA-SR**: This is the risk-sensitive algorithm that considers the Sharpe ratio based risk measure and updates the actor according to (44).

The underlying policy that guides the selection of the sign configuration in each of the algorithms above is a parameterized Boltzmann family and has the form

$$\mu_\theta(x, a) = \frac{e^{\theta^\top \phi_{x,a}}}{\sum_{a' \in A(x)} e^{\theta^\top \phi_{x,a'}}}, \quad \forall x \in \mathcal{X} \ , \ \forall a \in \mathcal{A}. \tag{45}$$

The simulations are conducted for 10000 time steps for all the algorithms. The road network used for conducting the experiments is shown in Fig. 3. Traffic is added to the network at each time step from the edge nodes, i.e., the nodes labelled **E** in Fig. 3. The spawn frequencies specify the rate at which traffic is generated at each edge node and follow the Poisson distribution. The spawn frequencies are set such that the proportion of number of vehicles on the main roads (the horizontal ones in Fig. 3) to those on the side roads is in the ratio $100 : 5$. This setting is close to what is observed in practice and has also been used for instance in Prashanth & Bhatnagar (2011; 2012). In all our experimetns, we set the weights in the single stage cost function (29) as follows: $r_1 = r_2 = 0.5$ and $r_2 = 0.6, s_2 = 0.4$. For the SPSA and SF based algorithms in the discounted setting, we set the parameter $\delta = 0.5$ and the discount factor $\gamma = 0.9$. The parameter α in the formulations (3) and (17) was set to 20. The step-size sequences are chosen as follows:

$$a(t) = \frac{1}{t}, \quad b(t) = \frac{1}{t^{0.75}}, \quad c(t) = \frac{1}{t^{0.66}}, t \geq 1. \tag{46}$$

Further, the constant k related to $d(t)$ is set to 1. It is easy to see that the choice of step-sizes above satisfies (A4). The results presented here are the averages over ten independent simulations with different initial seeds.

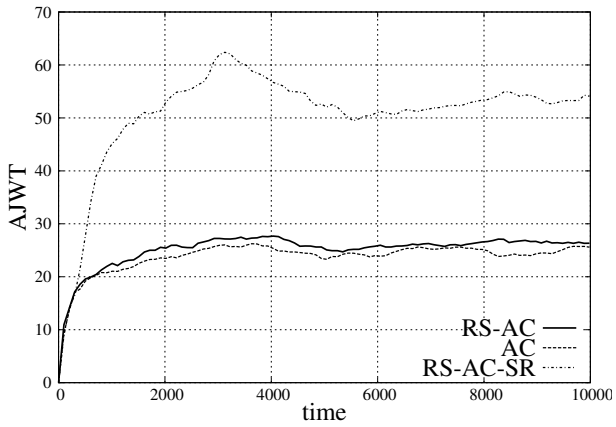


Figure 6. Performance Comparison of a plain AC algorithm vs. RS-AC algorithm using the average junction waiting time (AJWT)

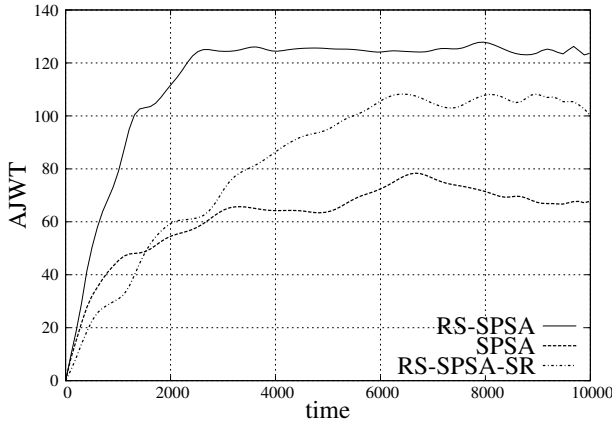


Figure 7. Performance Comparison of a SPSA based algorithm vs. RS-SPSA and RS-SPSA-SR algorithm using the average junction waiting time (AJWT)

We notice from the average junction waiting time plots in Figs. 5 and 7 that the performance of the risk sensitive variants RS-AC, RS-SPSA and RS-SF is very close to that of the algorithms AC, SPSA and SF, respectively. Further, we observe that the performance of the risk-sensitive algorithms that consider the Sharpe ratio based risk measure was significantly lesser than their counterparts that attempted to solve a constrained problem (3) and (17).