Policy Gradient Stochastic Approximation Algorithms for Adaptive Control of Constrained Time Varying Markov Decision Processes

Felisa J. Vázquez Abad
Département d’informatique et recherche opérationnelle
Université de Montréal, Québec H3C 3J7, Canada
email: vazquez@iro.umontreal.ca

Vikram Krishnamurthy,
Department of Electrical Engineering
University of British Columbia
(and University of Melbourne),
Vancouver, V6T1Z4, Canada
email: vikramk@ece.ubc.ca

Abstract—We present constrained stochastic approximation algorithms for computing the locally optimal policy of a constrained average cost finite state Markov Decision process. The stochastic approximation algorithms require computation of the gradient of the cost function with respect to the parameter that characterizes the randomized policy. This is computed by novel simulation based gradient estimation schemes involving weak derivatives. The algorithms proposed are simulation based and do not require explicit knowledge of the underlying parameters such as transition probabilities. We present three classes of algorithms based on primal dual methods, augmented Lagrangian (multiplier) methods and gradient projection primal methods. Unlike neuro-dynamic programming methods such as Q-Learning, the algorithms proposed here can handle constraints and time varying parameters.

I. INTRODUCTION

This paper presents stochastic approximation based algorithms for computing the optimal policy of a constrained average cost finite-state Markov Decision Process (MDP). Because the optimal control strategy of such problems is known to be a stationary randomized policy, in this work we focus on randomized policies to pose an optimization problem for the solution of the MDP. The problem can be stated in general terms in any convenient parameterization of the action probabilities, so that the optimization variable is a vector-valued parameter. The stochastic approximation algorithms require computation of the gradient of the cost function with respect to the parameter that characterizes the class of randomized policies. This is computed by novel simulation based gradient estimation schemes involving weak derivatives. In this paper we use a robust version of the weak derivative estimators that does not require knowledge of the underlying distribution. Similar to neuro-dynamic programming algorithms [6] (e.g., Q-learning and Temporal Difference methods), the algorithms proposed in this paper are simulation based and do not require explicit knowledge of the underlying parameters of the MDP such as transition probabilities (or equivalently invariant distributions). However, unlike Q-learning or Temporal Difference methods, the algorithms proposed here straightforwardly can handle constraints and time varying parameters.

A. The canonical control problem

Consider a unichain Markov Decision Process (MDP) \( \{X_n\}, \ n = 0, 1, \ldots \) with finite state space \( S \). When the system is in state \( i \in S \), a finite number of possible actions in the set \( A(i) \) can be taken, where \( A = \bigcup_{i \in S} A(i) \) is assumed to be finite. The evolution of the system is Markovian, and the kernel \( A(u) \) depends on the action \( u \in A(i) \), that is:

\[
P[X_{n+1} = j | X_n = i, u_n = u] = A_{ij}(u), \quad u \in A(i).
\]

Let \( \mathcal{F}_n = \sigma(X_1, \ldots, X_n, u_1, \ldots, u_{n-1}) \). The admissible policies are

\[
\mathcal{U} = \{ u = \{ u_n \} : u_n \text{ is measurable w.r.t. } \mathcal{F}_n, \forall n \in \mathbb{N} \}.
\]

The cost incurred at stage \( n \) is given by a known bounded function \( c(X_n, u_n) \geq 0 \) where \( c : S \times A \rightarrow \mathbb{R} \). For any admissible policy \( u \in \mathcal{U} \) define the average cost

\[
J_{x_0}(u) = \lim_{N \to \infty} \sup \frac{1}{N} \mathbb{E}_u \left[ \sum_{n=1}^{N} c(X_n, u_n) \mid X_0 = x_0 \right].
\]

The aim is to compute the optimal policy \( u^* \in \mathcal{U} \) s.t.

\[
J_{x_0}(u^*) = \inf_{u \in \mathcal{U}} J_{x_0}(u) \quad \forall x_0 \in S
\]

i.e., \( u^* \) has the minimum cost for all initial states \( x_0 \in S \).

Denote by \( \mathcal{U}_0 \) the class of randomized stationary Markovian policies, comprising all elements \( u \) that are stochastic processes (adapted to \( \mathcal{F}_n \)) of the form: \( \mathcal{U}_0 = \{ u = (u_n, n \in \mathbb{N}) : \forall n, u_n \sim \mu(X_n), \mu : x \in \mathcal{M}(A(x)), \forall x \in S \} \) where for any set \( G, \mathcal{M}(G) \) denotes the set of probability measures over \( G \). In other words, a randomized stationary policy is completely determined by a (conditional) distribution \( \mu(x) \) on \( A(x), x \in S \). By the unichain assumption, the limit in (2) exists for any \( x_0 \in S \) and \( u \in \mathcal{U}_0 \), and the limit is independent of \( x_0 \).

Motivated by several problems in telecommunication network optimization, we consider the MDP (3) subject to \( L \) average constraints of the form

\[
\lim_{N \to \infty} \frac{1}{N} \mathbb{E}_u \left[ \sum_{n=1}^{N} \beta_l(X_n, u_n) \right] \leq \gamma_l, \quad l = 1, \ldots, L
\]
where \( \beta_t : S \times A \to \mathbb{R} \) are known bounded functions and \( \gamma_t \) are known constants. It is well known that if \( L > 0 \) then the optimal policy \( u^* \in U \). Indeed, the optimal policy is randomized for at most \( L \) of the states.

We consider a “simulation” based system. By simulation based we mean that although the transition probabilities \( A(u), u \in A(i), i \in S \) of the underlying system are unknown, the system can be observed under any choice of control actions \( u = \{u_n\} \). In other words, the on-line learning methods that we study must be adapted to the compact subset of the Euclidean space.

Here \( a \) is the unknown, the system can be observed under any choice of control actions \( u = \{u_n\} \). In other words, the on-line learning methods that we study must be adapted to the compact subset of the Euclidean space.

**B. Parameterized Randomized Policies**

Define the action probabilities \( \theta \) parameterized by \( \psi \) as:

\[
\mathbb{P}[u_n = a|X_n = i] = \theta_{ia}(\psi), \quad a \in A(i), \quad i \in S
\]

where \( \sum_{a \in A(i)} \theta_{ia}(\psi) = 1 \) for every state \( i \in S \).

Here \( \psi \in \Psi \) is a finite dimensional vector which parameterizes the action probabilities \( \theta \). \( \Psi \) is some suitably defined compact subset of the Euclidean space.

An obvious parameterization of the action probabilities \( \theta \) is to choose \( \psi = \theta \). Indeed the parameterization used in [4] reduces to this. However, as shown in [2], [1] parameterization using spherical coordinates result in improved convergence properties. For this reason throughout this paper we deal with the following spherical coordinate parameterization.

Suppose without loss of generality that \( A(i) = \{0, \ldots, d(i)\} \). A novel solution was introduced in [8] for Hidden Markov Model parameter estimation through the use of spherical coordinates. Adapted to our MDP problem, it reads as follows: Fix the control agent \( i \in S \). To each value \( \theta_{ia}, a \in A(i) \) associate the values \( \lambda_{ia} = \sqrt{\theta_{ia}}. \) Then (5) yields \( \sum_{a \in A(i)} \lambda_{ia}^2 = 1 \) and \( \lambda_{ia} \) can be interpreted as the coordinates of a vector that lies on the surface of the unit sphere in \( \mathbb{R}^{d(i)+1} \), where \( d(i) + 1 \) is the size of \( A(i) \). In spherical coordinates, the angles are \( \alpha_{ip}, p = 1, \ldots, d(i), \) and the radius is always of size unity. If, for instance, \( d(i) = 2 \) (i.e., there are 3 actions), then

\[
\lambda_{i0} = \cos(\alpha_{i1}), \lambda_{i1} = \sin(\alpha_{i1}) \cos(\alpha_{i2}), \text{ and } \lambda_{i2} = \sin(\alpha_{i1}) \sin(\alpha_{i2})
\]

(because the radius is one). For the general case where \( d(i) \geq 2 \) in \( A(i) = \{0, \ldots, d(i)\} \), the spherical coordinates parameterization \( \alpha \) reads as follows: \( \lambda_{ia} = \sqrt{\theta_{ia}} \) and

\[
\lambda_{ia} = \begin{cases} 
\cos(\alpha_{i1}) & \text{if } a = 0 \\
\cos(\alpha_{i(a+1)}) \prod_{k=1}^{a} \sin(\alpha_{ik}) & 1 \leq a \leq d(i) - 1 \\
\sin(\alpha_{i(d(i))}) \prod_{k=1}^{d(i)-1} \sin(\alpha_{ik}) & a = d(i) 
\end{cases}
\]

(6)

Thus \( \alpha_{ip} \in \alpha \) where \( \alpha \) denotes the compact set

\[
\alpha = \{\alpha_{ip} \mod 2\pi \in [0, 2\pi]; \; i \in S, \; p \in P(i)\}
\]

(8)

Under this \( \alpha \) parameterization, the control variables \( \alpha_{ip}, p \in P(i) = \{1, \ldots, d(i)\} \) do not need to satisfy any physical constraints, since \( \sin(\alpha_{ip}) = \sin(\alpha_{ip} \mod 2\pi) \), etc.

**C. Parameterized Constrained MDP**

Consider the homogeneous Markov chain \( Z_n \) with state space in \( Z = S \times A \) and transition probabilities parameterized by \( \psi \) given by \( P_{a,i,j}^\psi(v) = P(X_{n+1} = j, u_{n+1} = a' \mid X_n = i, u_n = a) = \theta_{j,a}(\psi) A_j(a), i, j \in S, a \in A(i), a' \in A(j). \) From the unichain assumption it follows that for any \( \psi \in \Psi \) the chain \( \{Z_n\} \) is ergodic, and it possesses a unique invariant probability measure \( \pi_{i,a}(\psi); i \in S, a \in A(i) \).

Let \( E_\pi(\psi) \) denote expectation w.r.t measure \( \pi(\psi) \) parameterized by \( \psi \). From (2) we have

\[
J_\pi(u) = E_\pi(\psi)[c(Z)] \leq \sum_{i \in S} \sum_{a \in A(i)} \pi_{i,a}(\psi) c(i, a)
\]

The \( l = 1, \ldots, L \) constraints (4) can be expressed as

\[
B_l(\psi) = E_\pi(\psi)[\beta_i(Z)] - \gamma_l = \sum_{i \in S} \sum_{a \in A(i)} \pi_{i,a}(\psi) \beta_i(a) - \gamma_l \leq 0
\]

(9)

Thus the optimization problem (3) with constraints (4) can be written as

\[
\min_{\psi \in \Psi} C(\psi) \quad \text{subject to:} \quad B_l(\psi) \leq 0, \quad l = 1, \ldots, L, \quad (10)
\]

\[
\psi \in \Psi
\]

(11)

**Assumption 1:** For \( \psi \in \Psi, C(\psi), B(\psi) \in C^2 \) (twice continuously differentiable).

Indeed for \( \psi = \alpha \) or \( \theta, C(\cdot) \) and \( B(\cdot) \) are analytic functions (infinitely differentiable) in our signal model.

**Assumption 2:** The minima \( \psi^* \) of (10), (11), (12) are regular, i.e., \( \nabla_\psi B_l(\psi^*) \), \( l = 1, \ldots, L \) are linearly independent. Then \( \psi^* \) belongs to the set of Kuhn Tucker points

\[
K_T = \left\{ \psi^* \in \Psi : \exists \mu \geq 0, l = 1, \ldots, L \text{ such that} \quad \nabla_\psi C + \nabla_\psi B_l \mu = 0, \quad B_l' \mu = 0 \right\}
\]

(13)

where \( \mu = (\mu_1, \ldots, \mu_L)' \). Moreover \( \psi^* \) satisfies the second order sufficiency condition \( \nabla_\psi C(\psi^*) + \nabla_\psi B(\psi^*) \mu > 0 \) (positive definite) on the subspace \( \{y \in \mathbb{R}^L : \nabla_\psi B_l(\psi^*) y = 0\} \) for all \( l : B_l(\psi^*) = 0, \mu_l > 0 \).

2824
D. Summary of Main results

We use a gradient based stochastic approximation algorithm to optimize \( C(\psi) \) in (10) with respect to \( \theta(\psi) \) defined in (5). The stochastic approximation algorithm is of the form

\[
\psi(n+1) = \Pi \left[ \psi(n) - \epsilon(n) \nabla\psi C(\psi(n)) \right],
\]

(suitably modified with Lagrange multipliers to handle the constraints). Here \( \nabla\psi C \) denotes an estimate of the gradient \( \nabla\psi C, \Pi[\cdot] \) denotes an appropriate projection to deal with the constraints (11), and (12) and \( \epsilon(n) > 0 \) denotes the step size. If the parameters of the MDP are time-invariant, then there exists a unique optimum \( \psi^* \) for the constrained MDP (10), (11), (12).

1. MDP Constraints: For the MDP constraints (11), because \( B(\psi) \) is not a priori known (since the invariant measure \( \pi(\psi) \) is not known), it is not possible to use the usual gradient projection methods in [9] for example. In Sec.II, ODE versions of the primal dual algorithms with a penalty function and augmented Lagrangian (multiplier) algorithms are presented for solving the constrained MDP together with a convergence analysis. We also present a gradient-based hard projection algorithm.

2. Measured Valued Derivatives and Parameter Free Gradient Estimation: Since this was the main contribution of our recent papers [2], we only briefly outline the main idea here. In our paper [2], gradient estimators based on measured valued differentiation (weak derivatives) we derived to implement the stochastic gradient algorithm (14). Such algorithms have recently been proposed in the DES literature - see [7]. In particular, a parameter free gradient estimator based on frozen phantoms was developed. By filtering these frozen phantoms we derived in [2] a parameter free consistent algorithm for estimating gradients of the cost without explicit knowledge of the parameters of the MDP and without requiring off-line simulations of the system. The algorithms have significantly lower variance compared to score function gradient estimators which can unbounded variance estimation in some important cases.

3. Adaptive Control of Constrained MDP with time varying parameters: Putting the parameter free gradient estimation algorithm into a stochastic approximation algorithm (14) results in a new simulation based algorithm for solving the constrained MDP. Moreover, in this paper we are concerned with the case where the MDP parameters (e.g., transition probabilities \( A \)) can slowly change with time and hence we chose \( \epsilon(n) \) as a constant (non-decreasing) step size to track the time varying parameter \( \psi \). In Sec.IV, details of constant step size stochastic approximation algorithms based on the ODEs of Sec.II for solving the constrained MDP are presented. In particular, stochastic approximation versions of the primal dual, multiplier and projected gradient are presented. Weak convergence of the estimates \( \psi(n) \) to \( \psi^* \) is established for the constant step size algorithm (14).

E. Other approaches

The closest approach to that considered in this paper is the one presented in [4]. The MDP in [4] is identical to that considered here, but without the MDP constraints (11). [4] give an example of a parameterized class of policies which is identical to the coordinates \( \theta \). The approach for derivative estimation in [4] is via the Score Function method, which usually suffers from unbounded variance for infinite horizon costs. To alleviate this problem the authors use a forgetting factor that introduces a bias in the derivative estimation. Our derivative estimators are more efficient and consistent, with provably bounded variance. Typically, the score function method has a variance that is several orders of magnitude larger than the measured valued derivative estimator.

[4] uses a “simulation optimization” approach, where almost sure convergence to the true optimal value can be shown under an appropriate choice of the parameters of the algorithms. In particular all stochastic approximations involved in the above mentioned methodologies use decreasing step size. One of the motivations of the present work is to implement a stochastic approximation procedure with constant step size in order for the controlled Markov chain to be able to deal with tracking slowly varying external conditions which result in slowly varying \( A(u) \).

II. Deterministic Algorithms for Constrained MDP

In this section we consider the MDP problem (10) with MDP constraints (11) and parameter constraints (12). The corresponding stochastic approximation algorithms for the constrained MDP are given by replacing \( \nabla\alpha C, \nabla\alpha B, B \) in the deterministic algorithms presented below with the estimated gradients \( \tilde{\nabla}\alpha C, \tilde{\nabla}\alpha B, \tilde{B} \). These estimates are computed using the parameter free gradient estimation algorithms based on measured valued differentiation [2] – details omitted. The proofs of convergence of the resulting stochastic approximation algorithms are given in Sec.IV.

In Sec.II-A we present a penalty function based primal dual algorithm. In Sec.II-B we present an augmented Lagrangian algorithm. Both methods perform extremely well in numerical examples. For completeness we also present in Sec.II-C a primal algorithm based on gradient projection. This primal algorithm requires higher computational cost.

A. First-Order Primal Dual Algorithm

A widely used deterministic optimization method (with extension to stochastic approximation in [9, pg.180]) for handling constraints is based on the Lagrange multipliers and uses a first-order primal dual algorithm [5, pg 446]. First convert the inequality MDP constraints (11) to equality constraints by introducing the variables \( z = (z_1, \ldots, z_L) \) so
that $B_l(\alpha) + z_t^2 = 0$, $l = 1, \ldots, L$. Define $\psi \overset{\triangle}{=} (\alpha, z)$, $B_l(\psi) \overset{\triangle}{=} B_l(\alpha) + z_t^2$. Define the Lagrangian

$$L(\psi, \lambda) \overset{\triangle}{=} C(\alpha) + \sum_{l=1}^L \lambda_l B_l(\psi)$$

(15)

In order to converge, the primal dual algorithm requires the Lagrangian to be locally convex at the the optimum, i.e., Hessian to be positive definite at the optimum (which is much more restrictive that the second order sufficiency condition of Assumption 2 in Sec. I-C). We “convexify” the problem by adding a penalty term to (10). The resulting problem is:

$$\min_{\psi \in \Psi} C(\alpha) + \frac{\rho}{2} \sum_{l=1}^L (B_l(\psi))^2$$

subject to (11), (12). Here $\rho$ denotes a large positive constant.

As shown in [10, pg.429], the optimum of the above problem is identical to that of (10), (11), (12). Define the augmented Lagrangian,

$$L_\rho(\psi, \lambda) \overset{\triangle}{=} C(\alpha) + \sum_{l=1}^L \lambda_l B_l(\psi) + \frac{\rho}{2} \sum_{l=1}^L (B_l(\psi))^2$$

(16)

Note that although the original Lagrangian may not be convex near the solution (and hence the primal dual algorithm does not work), for sufficiently large $\rho$, the last term in $L_\rho$ “convexifies” the Lagrangian. Indeed, for sufficiently large $\rho$, [5] shows that the augmented Lagrangian, $L_\rho(\psi(n), \lambda(n))$, reads:

$$\alpha^{(n+1)} = \left[ \alpha^{(n)} - \epsilon \left( \nabla_\alpha C(\alpha^{(n)}) + \nabla_\alpha B(\alpha^{(n)}) \right) \right] \mod 2\pi$$

$$\lambda^{(n+1)} = \max \left[ 0, \lambda^{(n)} + \epsilon B(\alpha^{(n)}) \right]$$

(17)

(18)

where $\epsilon > 0$ denotes the step size.

**Lemma 1:** Under Assumptions 1 and 2 for sufficiently large $\rho > 0$, there exists $\tilde{\epsilon} > 0$ such that for all $\epsilon \in (0, \tilde{\epsilon})$, the sequence $\{\alpha^{(n)}, \lambda^{(n)}\}$ generated by the primal dual algorithm (17) is attracted to a local KT pair $(\alpha^*, \lambda^*)$.

Let $T \in \mathbb{R}^+$ denote a fixed constant and $t \in [0, T]$ denote the continuous time. Define the piecewise constant interpolated continuous-time process

$$\alpha^t(\epsilon) = \alpha^{(n)} \quad t \in [\epsilon n, (\epsilon + 1)n)$$

$$\lambda^t(\epsilon) = \lambda^{(n)} \quad t \in [\epsilon n, (\epsilon + 1)n)$$

(19)

(20)

**Theorem 1:** Under Assumptions 1 and 2, the interpolated process $(\alpha^t(\epsilon), \lambda^t(\epsilon))$ defined in (19), (20) converges uniformly as $\epsilon \to 0$ to the process $(\alpha(t), \lambda(t))$, i.e.,

$$\lim_{\epsilon \to 0} \limsup_{0 < t \leq T} |\alpha^t(\epsilon) - \alpha(t)| = 0, \quad \lim_{\epsilon \to 0} \limsup_{0 < t \leq T} |\lambda^t(\epsilon) - \lambda(t)| = 0$$

(21)

where $\{\alpha(t), \lambda(t)\}$ satisfy the ODEs

$$\frac{d}{dt} \alpha(t) = -\nabla_\alpha C(\alpha(t)) - \nabla_\alpha B(\alpha(t)) \lambda(t) + \rho B(\alpha(t))$$

$$\frac{d}{dt} \lambda(t) = \begin{cases} B_l(\alpha(t)) & \text{if } \lambda_l(t) \geq 0 \\ 0 & \text{if } \lambda_l(t) < 0 \\frac{\rho}{2} \lambda_l(t) < 0 \end{cases}$$

(22)

The attraction point of this ODE is the local KT pair $(\alpha^*, \lambda^*)$.

**B. Augmented Lagrangian (Multiplier) Algorithms**

1. **Inexact Primal Minimization Multiplier Algorithm:** The augmented Lagrangian approach (also known as a multiplier method) consists of the following coupled ODE and difference equation:

$$\frac{d\alpha^{(n+1)}(t)}{dt} = -\nabla_\alpha C(\alpha^{(n+1)}(t)) - \nabla_\alpha B(\alpha^{(n+1)}(t))$$

$$\left( \lambda(n) + \rho B(\alpha^{(n+1)}(t)) \right)$$

$$\lambda_l(n + 1) = \max \left[ 0, \lambda_l(n) + \rho B_l(\alpha^{(n+1)}(\infty)) \right]$$

(23)

(24)

where $\alpha^{(n+1)}(\infty)$ denotes the fixed point of the ODE (23). Iteration (24) is a first order update for the multiplier, while (23) represents an ODE which is attracted to the minimum of the augmented Lagrangian $L_\rho$. The max in (24) arises in dealing with the inequality constraints, see [5, pp.396]. [5, Proposition 4.2.3] shows that if $(\alpha^{(0)}(0), \lambda_0(0))$ lies in the domain of attraction of a local KT pair $(\alpha^*, \lambda^*)$, then (23), (24) converges to this KT pair.

A practical alternative to the above exact primal minimization is first order inexact minimization of the primal. The iterative version of the algorithm reads [5, pp.406]: At time $n + 1$ set $\alpha^{(0)}(n + 1) = \alpha(n)$. Then run $j = 0, \ldots, J - 1$ iterations of the following gradient minimization of the primal

$$\alpha^{(j+1)}(n + 1) = \left[ \alpha^{(j)}(n + 1) - \epsilon \left( \nabla_\alpha C(\alpha^{(j)}(n)) + \nabla_\alpha B(\alpha^{(j)}(n)) \right) \right] \mod 2\pi,$$

(25)

$$\alpha(n + 1) = \alpha^{(j)}(n + 1) \text{ followed by a first order multiplier step for } l = 1, \ldots, L$$

$$\lambda_l(n + 1) = \max \left[ 0, \lambda_l(n) + \rho B_l(\alpha(n + 1)) \right]$$

(26)

Iteration (25) represents a first order fixed step size inexact minimization of the augmented Lagrangian $L_\rho$ because (25) is terminated after a finite number of steps $J$, uces to (25), (26)

3. **Fixed Multiplier:** Fix $\lambda(n) = \bar{\lambda}$ for all time $n$ and only update $\alpha$ according to (25) with $J = 1$ iteration at each time instant. From Theorem 1, the interpolated trajectory of this
algorithm converges uniformly as $\epsilon \to 0$ to the trajectory of the ODE
\[
\frac{d}{dt} \alpha(t) = -\nabla\alpha C(\alpha(t)) - \nabla\alpha B(\alpha(t))(\bar{\lambda} + \rho B(\alpha(t))) \tag{27}
\]
First convert the $L$ inequality constraints to equality constraints $B_l(\psi) = \frac{\bar{\lambda}}{\rho} + \rho B(\alpha(t)) \in \mathbb{R}^{L+1}$ as defined in Sec.II-A. Let $\psi^*, \lambda^*$ denote the corresponding KT pair where $\psi^* = (\alpha^*, z^*)$.

**Result 1:** [5, Proposition 4.2.3]. Let $\rho > 0$ be scalar such that $\nabla\psi^2 L(\psi^*, \lambda^*) > 0$. Then there exist positive scalars $\delta$ and $K$ such that for $(\bar{\lambda}, \rho) \in D \in \mathbb{R}^{L+1}$ defined by
\[
D = \{(\lambda, \rho) : \|\lambda - \lambda^*\| < \delta \rho, \ \rho \geq \bar{\rho}\}
\]
the attraction point of the ODE $\alpha^\lambda, \rho$ is unique. Moreover
\[
\|\alpha^\lambda, \rho - \alpha^*\| \leq K \frac{\|\bar{\lambda} - \lambda^*\|}{\rho}
\]

**C. Projected Gradient Primal Algorithm**

In this subsection, we present a “ primal algorithm” [10] for solving the constrained MDP problem. The algorithm is a gradient projection method (see [5]) that updates along the gradient direction and then projects the resulting vector onto the constraints surface (if the resulting vector is infeasible).

\[
\tilde{\alpha}(n+1) = \alpha(n) - \epsilon \nabla\alpha C(\alpha(n)) \tag{28}
\]
\[
\alpha(n+1) = \Pi_X(\tilde{\alpha}(n+1)),
\]
where $\Pi_X(\cdot)$ is the projection of the vector onto the space:
\[
X = \{\alpha \in \mathbb{R}^{d-1} : B_l(\alpha) \leq 0, \ l = 1, \ldots, L\}.
\]

Evaluation of the projection operation is the main computational hindrance of the algorithm. Observe that if $x = \Pi_X(c)$, then $x$ is the vector in $X$ that minimizes the distance $\|x - c\|^2$, or equivalently, $x$ is the solution to the quadratic optimization problem:
\[
\min_{x \in X} \left( \frac{1}{2} x' x - c' x \right).
\]
Because the points $\alpha(n), \tilde{\alpha}(n+1)$ are close when $\epsilon$ is small, using the fact that the constraints are analytical functions of $\alpha$ (polynomial and trigonometric functions are analytic), a Taylor approximation of $B(x)$ can be used to obtain the following (approximative) optimization problem with linear constraints (c.f. Example 3.4.3 of [5]):
\[
\min_{x \in X} \left( \frac{1}{2} x' x - c' x \right),
\]
subject to:
\[
B(\alpha(n)) + \nabla\alpha B(\alpha(n))'(x - \alpha(n)) \leq 0.
\]
Let $\mu = (\mu_1, \ldots, \mu_L)$ be the dual variable. Then the dual to this problem is:
\[
\min_{\mu \geq 0} \left( \frac{1}{2} \mu' \nabla\alpha B(\alpha(n))'/\nabla\alpha B(\alpha(n)) \mu - [B(\alpha(n)) + \nabla\alpha B(\alpha(n))'(\tilde{\alpha}(n+1) - \alpha(n))]' \mu \right),
\]
and the optimal solutions of the dual and primal problems satisfy $x^* = \tilde{\alpha}(n+1) - \nabla\alpha B(\alpha(n))' \mu^* \approx \Pi_X(\tilde{\alpha}(n+1))$.

Although a linearization has been used to find the projection, as $\epsilon \to 0$, one would hope that this procedure converges to the optimal feasible point. The subsidiary optimization problem must be solved at each iteration of the procedure (28).

**III. Measure-Valued Gradient Estimation and Learning Based Constrained MDP Solution**

We refer the reader to to the companion papers [2], [3] and also the submitted journal paper [1], on how a measure-valued gradient estimator can be implemented which does not require explicit knowledge of the parameters of the constrained MDP.

**IV. Stochastic Approximation Algorithms and Convergence**

In this section we present weak convergence proofs for the learning algorithms that use the parameter free gradient estimators of Sec.III. These learning algorithms are stochastic versions of the deterministic algorithms presented in Sec.II.

**Notation:** $E_{\pi(\psi)}$ denotes expectation w.r.t measure $\pi(\psi)$ parameterized by $\psi$; $E_{\phi}$ denotes expectation w.r.t. the underlying probability measure of $\{Z_1, \ldots, Z_N\}$, where $N \geq 1$ is a fixed positive integer. We only consider equality constraints here (as mentioned in Section II-B) the inequality constraints can be handled with minor modifications) so that the constrained MDP problem (10), (11), (12) reads
\[
\min_{\alpha_i \in \mathbb{R}^{d-1}, i \in S} \frac{C(\alpha)}{s.t. \ B_l(\alpha) = 0, \ l = 1, \ldots, L}
\]

A control “agent” is associated with each of the possible visited states $i \in S$. The control parameter for this agent is the vector $(\alpha_{i, \psi}, \rho \in \mathcal{P}(i))$, plus an agent for the (artificial control) variable representing the Lagrange multiplier $\lambda$. The scheme works by observing the process over an estimation interval, during which the value of the control parameter does not change and the estimators $\tilde{C}(n), B^t(n)$ for the sample averages are calculated, along with the gradient estimators of the form $\nabla C^t(n), \nabla B^t(n)$. The computation of each partial derivative can be made locally at each controller. Let $h_0(\alpha) = E_{\pi(\psi)}(\nabla C^t(n)), h_i(\alpha) = E_{\pi(\psi)}(\nabla B^t(n))$ be the invariant averages of the batch estimation.

1) **First-Order Primal Dual Algorithm:** Consider the stochastic approximation procedure where the control parameter $(\alpha, \lambda)$ is updated with (c.f. (17), (18))
\[
\alpha'(n+1) = \alpha'(n) - \epsilon \left( \nabla \tilde{C}(n) + \sum_{i=1}^L \lambda_i(n) + \rho B^t_i(n) \tilde{B}_i^t(n) \right) \tag{29}
\]
\[
\lambda'(n+1) = \lambda'(n) + \epsilon \tilde{B}^t(n) \tag{30}
\]
Proposition 1: Define the interpolated process \( \alpha'(t) \) as in (19) and analogously for \( \lambda'(t) \). Then as \( \epsilon \to 0 \), the interpolated process converges in distribution to the solution of the ODE (c.f. (22))

\[
\frac{d}{dt} \alpha(t) = - \left[ h_0[\alpha(t)] \right. \\
\left. + \sum_{l=1}^{L} (\tilde{\lambda}_l + \rho B_l[\alpha(t)]) h_l[\alpha(t)] + \kappa[\alpha(t)] \right] 
\]

(31)

\[
\frac{d}{dt} \lambda(t) = B[\alpha(t)],
\]

where the added drift is defined as:

\[
\kappa(\alpha) = \rho \lim_{n \to \infty} \sum_{l=1}^{L} \text{Cov}_{\epsilon, \psi}[\tilde{B}_l^\epsilon(n), \tilde{B}_l^\epsilon(n)].
\]

2) Augmented Lagrangian Multiplier Algorithm – two time scale algorithm: Consider the following stochastic approximation version of the multiplier algorithm (26):

\[
\alpha'(n + 1) = \alpha'(n) - \epsilon \left( \tilde{\nabla}C^\epsilon(n) \\
+ \sum_{l=1}^{L} (\tilde{\lambda}_l(n) + \rho B_l[\alpha'(n)]) \tilde{h}_l(n) \right).
\]

(32)

where \( \{\lambda'(n), \epsilon > 0, n \in \mathbb{N}\} \) is any tight sequence. A trivial example is when \( \lambda'(n) \) is a bounded constant \( \alpha(t) \).

Proposition 2: Assume that \( \{\lambda'(n), \epsilon > 0, n \in \mathbb{N}\} \) is tight. Define the interpolated process \( \alpha'(t) \) of (32) as in (19). Then as \( \epsilon \to 0 \), the interpolated process \( \alpha'(t) \) converges in distribution to the solution of the ODE:

\[
\frac{d \alpha(t)}{dt} = - \left[ h_0[\alpha(t)] + \sum_{l=1}^{L} (\tilde{\lambda}_l + \rho B_l[\alpha(t)]) h_l[\alpha(t)] + \kappa[\alpha(t)] \right],
\]

(33)

where \( \tilde{\lambda}_l \) is an accumulation point of the sequence \( \{\tilde{\lambda}(\epsilon), \epsilon > 0\} \) of (convergent) Cesaro sums:

\[
\tilde{\lambda}(\epsilon) \equiv \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \lambda'(n),
\]

and the added drift is defined as:

\[
\kappa(\alpha) = \rho \lim_{n \to \infty} \sum_{l=1}^{L} \text{Cov}_{\epsilon, \psi}[\tilde{B}_l^\epsilon(n), \tilde{B}_l^\epsilon(n)].
\]

Consider the following update of the multiplier \( \lambda(\epsilon) \) in (32). Define \( J = [1/\epsilon] \) and consider the recursion

\[
\lambda(n + 1) = \lambda(n) + \tilde{B}(n/J) 1_{\{\epsilon \in \mathbb{N}\}}
\]

(34)

together with (32). Thus the multiplier is updated once every \( J \) time points. Here

\[
\tilde{B}(n/J) = \frac{1}{J} \sum_{j=\lfloor (n-1)J \rfloor + 1}^{nJ} \hat{B}(j)
\]

Then as \( \epsilon \to 0 \), the algorithm (34), (32) converges weakly to the deterministic system (23), (24) which is the exact multiplier algorithm. As mentioned in Sec.II-B this in turn converges to a local KT point. In a practical implementation, one would choose \( J \) as a large positive integer. In our numerical examples even a choice of \( J = 10 \) resulted in convergence to a KT point.

Remarks: We refer the reader to [1] for further details and proofs. In a companion paper [3], a detailed numerical study of the frozen phantoms (parameter free gradient estimators) is conducted. The effect of moving averages and exponentially discounted averages on the bias is also studied. See also [2] for further numerical examples. In current work, we are examining applications of the techniques in this paper to admission control of wireless networks.

V. REFERENCES