# Quasi Stochastic Approximation

American Control Conference
San Francisco, June 2011

Sean P. Meyn

Joint work with Darshan Shirodkar and Prashant Mehta

Coordinated Science Laboratory
and the Department of Electrical and Computer Engineering
University of Illinois at Urbana-Champaign, USA

Thanks to NSF & AFOSR

# Outline

# Background: Stochastic Approximation

## Robbins and Monro

### Stochastic Approximation

Setting: Solve the equation $\overline{h}(\vartheta) = 0$, with

$$\overline{h}(\vartheta) = \mathsf{E}[h(\vartheta, \zeta)]\,, \quad \text{where } \zeta \text{ is random.}$$

Robbins and Monro[5]: Fixed point iteration with noisy measurements,

$$\vartheta_{n+1} = \vartheta_n + a_n h(\vartheta_n, \zeta_n)\,, \qquad n \geq 0, \quad \vartheta_0 \in \mathbb{R}^d \text{ given.}$$

# Background: Stochastic Approximation

## Robbins and Monro

### Stochastic Approximation

Setting: Solve the equation $\overline{h}(\vartheta) = 0$, with

$$\overline{h}(\vartheta) = \mathsf{E}[h(\vartheta, \zeta)]\,, \quad \text{where } \zeta \text{ is random.}$$

Robbins and Monro[5]: Fixed point iteration with noisy measurements,

$$\vartheta_{n+1} = \vartheta_n + a_n h(\vartheta_n, \zeta_n)\,, \qquad n \geq 0, \quad \vartheta_0 \in \mathbb{R}^d \text{ given.}$$

Typical assumptions for convergence:

- Step size $a_n = (1+n)^{-1}$
- Random sequnce $\zeta_n$ is identical to $\zeta$ in distribution and i.i.d..
- Global stability of ODE $\dot{\theta} = \overline{h}(\theta)$.

Excellent recent reference: Borkar 2008[2].

# Background: Stochastic Approximation

## Variance

### Linearization of Stochastic Approximation

Assuming convergence, write $\tilde{\vartheta}_{n+1} \approx a_n \left[ A\tilde{\vartheta}_n + Z_n \right]$

with $Z_n = h(\vartheta^*, \zeta_n)$ (zero mean), and $A = \nabla \overline{h}\,(\vartheta^*)$.

# Background: Stochastic Approximation

## Variance

### Linearization of Stochastic Approximation

Assuming convergence, write $\tilde{\vartheta}_{n+1} \approx a_n \left[ A\tilde{\vartheta}_n + Z_n \right]$
with $Z_n = h(\vartheta^*, \zeta_n)$ (zero mean), and $A = \nabla \overline{h}\,(\vartheta^*)$.

Central Limit Theorem: $\sqrt{n}\tilde{\vartheta}_n \sim N(0, \Sigma_\vartheta)$

# Background: Stochastic Approximation

## Variance

### Linearization of Stochastic Approximation

Assuming convergence, write $\tilde{\vartheta}_{n+1} \approx a_n \left[ A\tilde{\vartheta}_n + Z_n \right]$
with $Z_n = h(\vartheta^*, \zeta_n)$ (zero mean), and $A = \nabla \overline{h}\,(\vartheta^*)$.

### Central Limit Theorem: $\sqrt{n}\tilde{\vartheta}_n \sim N(0, \Sigma_\vartheta)$

Holds under mild conditions.
Lyapunov equation for variance, *requires* $\operatorname{eig}(A) < -\frac{1}{2}$:

$$(A + \tfrac{1}{2}I)\Sigma_\vartheta + \Sigma_\vartheta (A + \tfrac{1}{2}I)^\intercal + \Sigma_Z = 0$$

# Background: Stochastic Approximation

## Stochastic Newton Raphson

### Corollary to CLT

*Question*: What is the optimal matrix gain $\Gamma_n$?

$$\vartheta_{n+1} = \vartheta_n + a_n \Gamma_n h(\vartheta_n, \zeta_n)\,, \qquad n \geq 0, \quad \vartheta_0 \in \mathbb{R}^d \text{ given.}$$

# Background: Stochastic Approximation

## Stochastic Newton Raphson

### Corollary to CLT

*Question*: What is the optimal matrix gain $\Gamma_n$?

$$\vartheta_{n+1} = \vartheta_n + a_n \Gamma_n h(\vartheta_n, \zeta_n)\,, \qquad n \geq 0, \quad \vartheta_0 \in \mathbb{R}^d \text{ given.}$$

*Answer*: Stochastic Newton-Raphson, so that $\Gamma_n^* \to -A^{-1}$.

# Background: Stochastic Approximation

## Stochastic Newton Raphson

### Corollary to CLT

*Question*: What is the optimal matrix gain $\Gamma_n$?

$$\vartheta_{n+1} = \vartheta_n + a_n \Gamma_n h(\vartheta_n, \zeta_n)\,, \qquad n \geq 0, \quad \vartheta_0 \in \mathbb{R}^d \text{ given.}$$

*Answer*: Stochastic Newton-Raphson, so that $\Gamma_n^* \to -A^{-1}$.

Proof: Linearization reduces to standard Monte-Carlo

# Example: Root finding

$h(\vartheta, \zeta) = 1 - \tan(\vartheta) + \zeta$. $\zeta$, normal random variable $N(0, 9)$

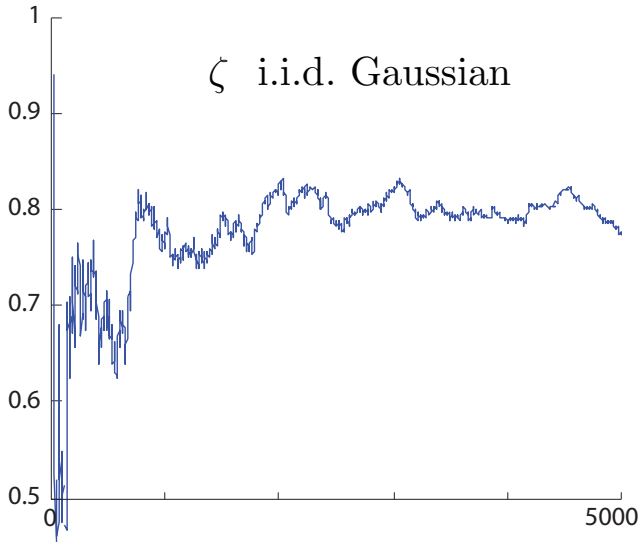
$$h(\vartheta, \zeta) = 1 - \tan(\vartheta) + \zeta$$
$$\overline{h}(\vartheta) = 1 - \tan(\vartheta) \qquad \Longrightarrow \quad \vartheta^* = \pi/4.$$

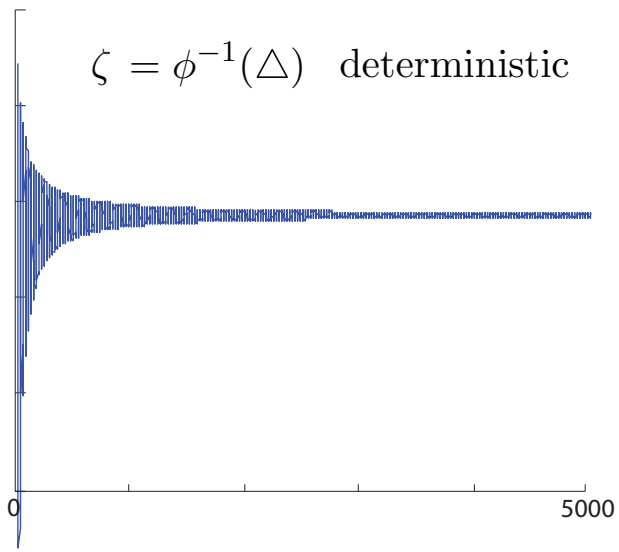
# Example: Root finding

$h(\vartheta, \zeta) = 1 - \tan(\vartheta) + \zeta$. $\zeta$, normal random variable $N(0, 9)$

$$h(\vartheta, \zeta) = 1 - \tan(\vartheta) + \zeta$$
$$\overline{h}(\vartheta) = 1 - \tan(\vartheta) \qquad \Longrightarrow \quad \vartheta^* = \pi/4.$$

Estimates of $\vartheta^*$ using i.i.d. and deterministic sequences (each $\sim \zeta$):



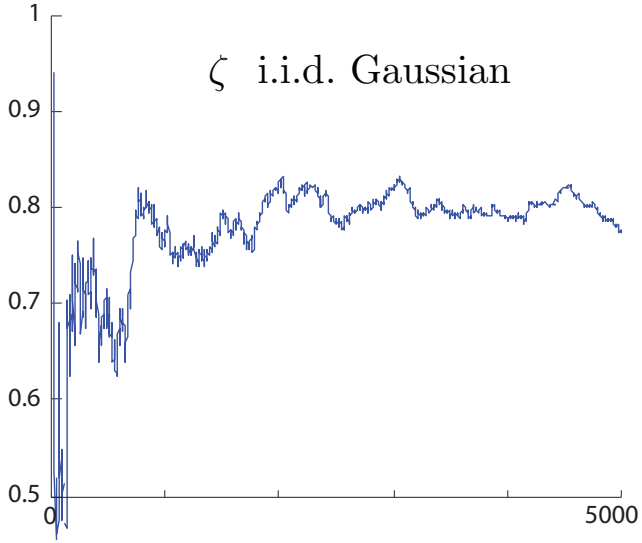



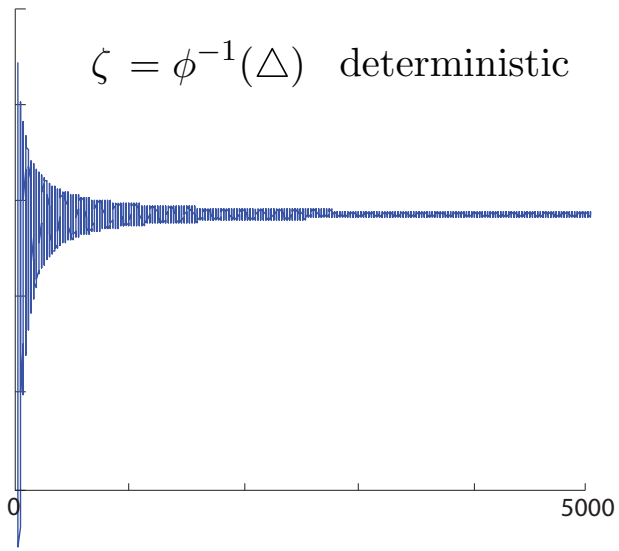
# Example: Root finding

$h(\vartheta, \zeta) = 1 - \tan(\vartheta) + \zeta$. $\zeta$, normal random variable $N(0, 9)$

$$h(\vartheta, \zeta) = 1 - \tan(\vartheta) + \zeta$$
$$\overline{h}(\vartheta) = 1 - \tan(\vartheta) \qquad \Longrightarrow \quad \vartheta^* = \pi/4.$$

Estimates of $\vartheta^*$ using i.i.d. and deterministic sequences (each $\sim \zeta$):




Analysis requires theory of quasi-stochastic approximation...

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x, u)$ cost: $c(x, u)$

HJB equation for discounted-cost optimal control problem,

$$\min_u \{\underbrace{c(x, u) + f(x, u) \cdot \nabla h\,(x)}\} = \gamma h(x)$$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

HJB equation for discounted-cost optimal control problem,

$$\min_u \{\underbrace{c(x,u) + f(x,u) \cdot \nabla h\,(x)}_{\text{Q function}}\} = \gamma h(x)$$

$$Q(x,u) = c(x,u) + f(x,u) \cdot \nabla h(x)$$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

HJB equation for discounted-cost optimal control problem,

$$\min_u \{ \underbrace{c(x,u) + f(x,u) \cdot \nabla h\,(x)}_{\text{Q function}} \} = \gamma h(x)$$

Fixed point equation for Q-function: Writing $\underline{Q} = \min_u Q(x,u) = \gamma h(x)$,

$$Q(x,u) = c(x,u) + f(x,u) \cdot \nabla h(x)$$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

HJB equation for discounted-cost optimal control problem,

$$\min_u \{\underbrace{c(x,u) + f(x,u) \cdot \nabla h\,(x)}_{\text{Q function}}\} = \gamma h(x)$$

Fixed point equation for Q-function: Writing $\underline{Q} = \min_u Q(x,u) = \gamma h(x)$,

$$Q(x,u) = c(x,u) + f(x,u) \cdot \nabla h(x) = c(x,u) + \gamma^{-1} f(x,u) \cdot \nabla \underline{Q}\,(x)$$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

HJB equation for discounted-cost optimal control problem,

$$\min_u \{ \underbrace{c(x,u) + f(x,u) \cdot \nabla h\,(x)}_{\text{Q function}} \} = \gamma h(x)$$

Fixed point equation for Q-function: Writing $\underline{Q} = \min_u Q(x,u) = \gamma h(x)$,

$$Q(x,u) = c(x,u) + f(x,u) \cdot \nabla h(x) = c(x,u) + \gamma^{-1} f(x,u) \cdot \nabla \underline{Q}\,(x)$$

Parameterization set of approximations, $\{Q^\vartheta(x,u) : \vartheta \in \mathbb{R}^d\}$

Bellman error: $\mathcal{E}^\vartheta(x,u) = \gamma(Q^\vartheta(x,u) - c(x,u)) - f(x,u) \cdot \nabla \underline{Q}^\theta\,(x)$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

Parameterization, $\{Q^\vartheta(x,u) : \vartheta \in \mathbb{R}^d\}$

Bellman error: $\mathcal{E}^\vartheta(x,u) = \gamma(Q^\vartheta(x,u) - c(x,u)) - f(x,u) \cdot \nabla \underline{Q}^\theta(x)$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

Parameterization, $\{Q^\vartheta(x,u) : \vartheta \in \mathbb{R}^d\}$

Bellman error: $\mathcal{E}^\vartheta(x,u) = \gamma(Q^\vartheta(x,u) - c(x,u)) - f(x,u) \cdot \nabla \underline{Q}^\theta(x)$

Model-free form:

$$\mathcal{E}^\vartheta(x,u) = \gamma(Q^\vartheta(x,u) - c(x,u)) - \tfrac{d}{dt}\underline{Q}^\theta(x(t))\Big|_{x=x(t),\ u=u(t)}$$

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Background: Q-learning

M&M 2009 system: $\dot{x} = f(x,u)$ cost: $c(x,u)$

## Parameterization, $\{Q^\vartheta(x,u) : \vartheta \in \mathbb{R}^d\}$

Bellman error: $\mathcal{E}^\vartheta(x,u) = \gamma(Q^\vartheta(x,u) - c(x,u)) - f(x,u)\cdot\nabla \underline{Q}^\theta\,(x)$

Model-free form:

$$\mathcal{E}^\vartheta(x,u) = \gamma(Q^\vartheta(x,u) - c(x,u)) - \tfrac{d}{dt}\underline{Q}^\theta\,(x(t))\Big|_{x=x(t),\ u=u(t)}$$

## Q-learning of M&M

- Find zeros of $\overline{h}(\vartheta) = \nabla \mathsf{E}[\mathcal{E}^\vartheta(x,u)^2]$
- $\zeta = (x_\infty, u_\infty)$ ergodic steady-state.
- Choose input: stable feedback + mixture of sinusoids, $u(t) = -k(x(t)) + \omega(t)$,

Watkins & Dayan, Q-learning, 1992; Bertsekas & Tsitsiklis, NDP, 1996; Mehta & Meyn, CDC 2009

# Example: Q-Learning

$\dot{x} = -x^3 + u, \;\; c(x,u) = \frac{1}{2}(x^2 + u^2)$

HJB Equation: $\min_u \left[ c(x,u) + (-x^3 + u) \cdot \nabla h\,(x) \right] = \gamma h(x)$

# Example: Q-Learning

$\dot{x} = -x^3 + u, \;\; c(x,u) = \frac{1}{2}(x^2 + u^2)$

HJB Equation: $\min_u \big[c(x,u) + (-x^3 + u) \cdot \nabla h\,(x)\big] = \gamma h(x)$

Basis: $Q^\vartheta(x,u) = c(x,u) + \vartheta_1 x^2 + \vartheta_2 \dfrac{xu}{1+2x^2}$

# Example: Q-Learning

$\dot{x} = -x^3 + u, \;\; c(x,u) = \frac{1}{2}(x^2 + u^2)$

HJB Equation: $\min_u \big[c(x,u) + (-x^3 + u) \cdot \nabla h\,(x)\big] = \gamma h(x)$

Basis: $Q^{\vartheta}(x,u) = c(x,u) + \vartheta_1 x^2 + \vartheta_2 \dfrac{xu}{1+2x^2}$

Control: $u(t) = A[\sin(t) + \sin(\pi t) + \sin(et)]$

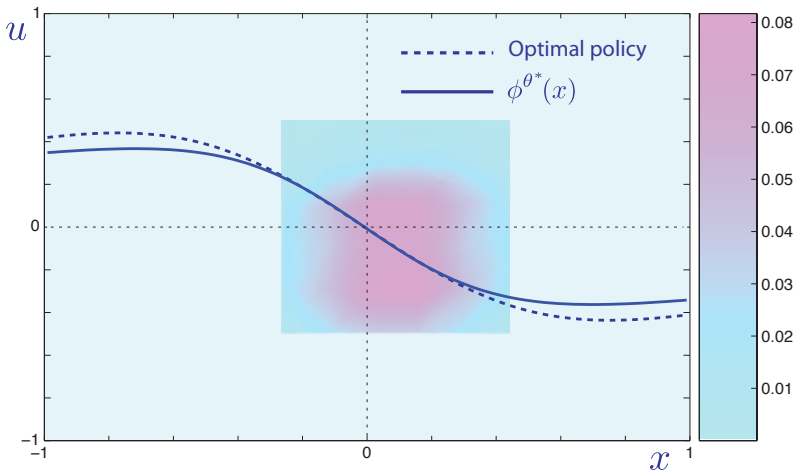
# Example: Q-Learning

$\dot{x} = -x^3 + u, \;\; c(x,u) = \frac{1}{2}(x^2 + u^2)$

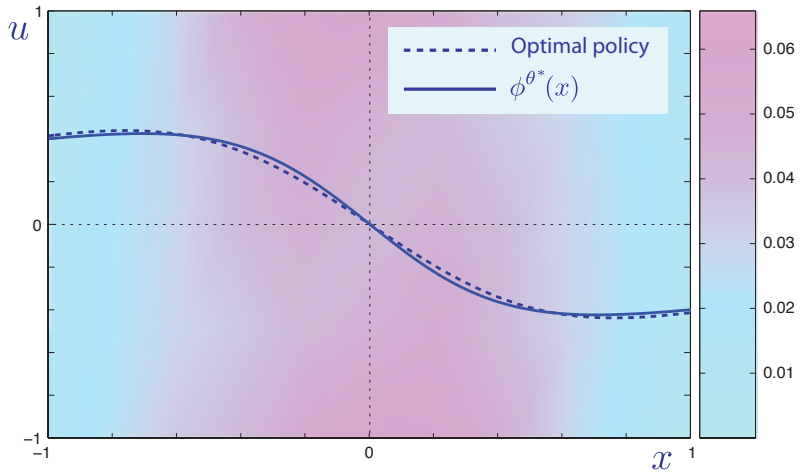
HJB Equation: $\min_u \big[c(x,u) + (-x^3 + u) \cdot \nabla h\,(x)\big] = \gamma h(x)$

Basis: $Q^\vartheta(x,u) = c(x,u) + \vartheta_1 x^2 + \vartheta_2 \dfrac{xu}{1+2x^2}$

Control: $u(t) = A[\sin(t) + \sin(\pi t) + \sin(et)]$




Low amplitude input




High amplitude input

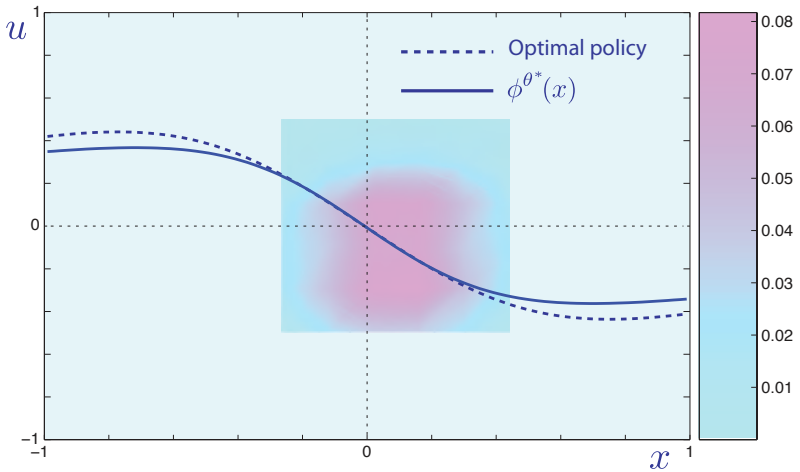
# Example: Q-Learning

$\dot{x} = -x^3 + u, \quad c(x,u) = \frac{1}{2}(x^2 + u^2)$

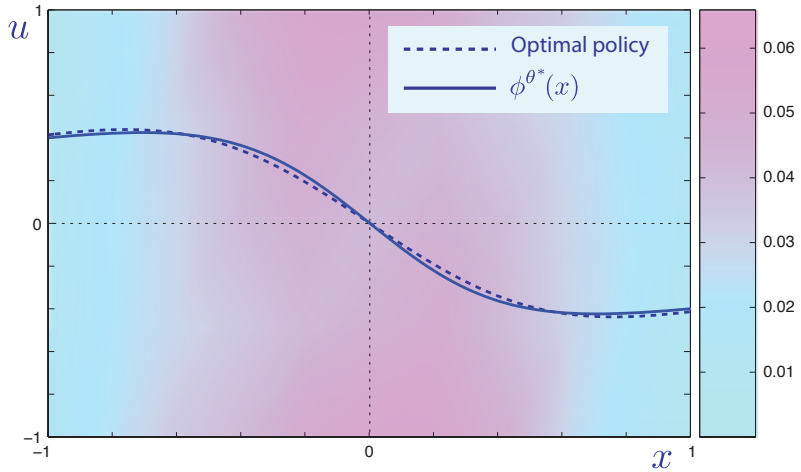
HJB Equation: $\min_u \big[c(x,u) + (-x^3 + u) \cdot \nabla h\,(x)\big] = \gamma h(x)$

Basis: $Q^\vartheta(x,u) = c(x,u) + \vartheta_1 x^2 + \vartheta_2 \dfrac{xu}{1+2x^2}$

Control: $u(t) = A[\sin(t) + \sin(\pi t) + \sin(et)]$




Low amplitude input




High amplitude input

Analysis requires theory of quasi-stochastic approximation...

# Quasi-Stochastic Approximation

## Continuous time, deterministic version of SA

$$\boxed{\frac{d}{dt}\vartheta(t) = a(t)h(\vartheta(t), \zeta(t))}$$

# Quasi-Stochastic Approximation

## Continuous time, deterministic version of SA

$$\boxed{\frac{d}{dt}\vartheta(t) = a(t)h(\vartheta(t), \zeta(t))}$$

Assumptions

- *Ergodicity*: $\boldsymbol{\zeta}$ satisfies

$$\overline{h}(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T h(\theta, \zeta(t))\, dt, \quad \text{for all } \theta \in \mathbb{R}^d.$$

- *Decreasing gain*: $a(t) \downarrow 0$, with

$$\int_0^\infty a(t)\, dt = \infty, \qquad \int_0^\infty a(t)^2\, dt < \infty$$

usually, take $a(t) = (1+t)^{-1}$

# Quasi-Stochastic Approximation

## Continuous time, deterministic version of SA

$$\boxed{\frac{d}{dt}\vartheta(t) = a(t)h(\vartheta(t), \zeta(t))}$$

Assumptions

- *Ergodicity*: $\boldsymbol{\zeta}$ satisfies

$$\overline{h}(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T h(\theta, \zeta(t))\, dt, \quad \text{for all } \theta \in \mathbb{R}^d.$$

- *Decreasing gain*: $a(t) \downarrow 0$, with

$$\int_0^\infty a(t)\, dt = \infty, \qquad \int_0^\infty a(t)^2\, dt < \infty$$

usually, take $a(t) = (1+t)^{-1}$

- *Stable ODE*: $\overline{h}$ is globally Lipschitz, and $\dot{\vartheta} = \overline{h}(\vartheta)$ is globally asymptotically stable, with globally Lipschitz Lyapunov function

# Quasi-Stochastic Approximation

## Stability & Convergence

$$\frac{d}{dt}\vartheta(t) = a(t)h(\vartheta(t), \zeta(t))$$

Assumptions

- *Ergodicity*: $\zeta$ satisfies

$$\overline{h}(\theta) = \lim_{T\to\infty} \frac{1}{T}\int_0^T h(\theta, \zeta(t))\, dt, \quad \text{for all } \theta \in \mathbb{R}^d.$$

- *Decreasing gain*: $a(t) \downarrow 0$, with

$$\int_0^\infty a(t)\, dt = \infty, \quad \int_0^\infty a(t)^2\, dt < \infty$$

- *Stable ODE*: $\overline{h}$ is globally Lipschitz, and $\dot{\vartheta} = \overline{h}(\vartheta)$ is globally asymptotically stable, with globally Lipschitz Lyapunov function

Theorem: $\vartheta(t) \to \vartheta^*$ for any initial condition.

# Quasi-Stochastic Approximation

## Variance

- $a(t) = 1/(1+t)$
- The model is linear: $h(\theta, \zeta) = A\theta + \zeta$, and each $\lambda(A)$ satisfies $\mathrm{Re}(\lambda) < -1$.
- $\boldsymbol{\zeta}$ has zero mean: $\lim_{T\to\infty} \frac{1}{T} \int_0^T \zeta(t)\, dt = 0$, at rate $1/T$.

# Quasi-Stochastic Approximation

## Variance

- $a(t) = 1/(1+t)$
- The model is linear: $h(\theta, \zeta) = A\theta + \zeta$, and each $\lambda(A)$ satisfies $\mathrm{Re}(\lambda) < -1$.
- $\boldsymbol{\zeta}$ has zero mean: $\lim_{T\to\infty} \frac{1}{T} \int_0^T \zeta(t)\, dt = 0$, at rate $1/T$.

Rate of convergence is $t^{-1}$, and not $t^{-\frac{1}{2}}$, under these assumptions:

# Quasi-Stochastic Approximation

## Variance

- $a(t) = 1/(1+t)$
- The model is linear: $h(\theta, \zeta) = A\theta + \zeta$, and each $\lambda(A)$ satisfies $\mathrm{Re}(\lambda) < -1$.
- $\boldsymbol{\zeta}$ has zero mean: $\lim_{T\to\infty} \frac{1}{T} \int_0^T \zeta(t)\, dt = 0$, at rate $1/T$.

Rate of convergence is $t^{-1}$, and not $t^{-\frac{1}{2}}$, under these assumptions:

Theorem: For some constant $\bar{\sigma} < \infty$,

$$\limsup_{t\to\infty} t\|\vartheta(t) - \vartheta^*\| \le \bar{\sigma}$$

# Polyak and Juditsky

Polyak and Juditsky[4, 2] obtain *optimal variance* for SA using a very simple approach: *High gain, and averaging*.

# Polyak and Juditsky

Polyak and Juditsky[4, 2] obtain *optimal variance* for SA using a very simple approach: *High gain, and averaging*.

Deterministic P&J: Choose *high-gain*, $a(t) = 1/(1+t)^{\delta}$, with $\delta \in (0, 1)$.

$$\frac{d}{dt}\gamma(t) = \frac{1}{(1+t)^{\delta}} h(\gamma(t), \zeta(t))$$

# Polyak and Juditsky

Polyak and Juditsky[4, 2] obtain *optimal variance* for SA using a very simple approach: *High gain, and averaging*.

Deterministic P&J: Choose *high-gain*, $a(t) = 1/(1+t)^\delta$, with $\delta \in (0, 1)$.

$$\frac{d}{dt}\gamma(t) = \frac{1}{(1+t)^\delta} h(\gamma(t), \zeta(t))$$

The output of this algorithm is then averaged:

$$\frac{d}{dt}\vartheta(t) = \frac{1}{1+t}\big(-\vartheta(t) + \gamma(t)\big)$$

# Polyak and Juditsky

Polyak and Juditsky[4, 2] obtain *optimal variance* for SA using a very simple approach: *High gain, and averaging*.

Deterministic P&J: Choose *high-gain*, $a(t) = 1/(1+t)^\delta$, with $\delta \in (0,1)$.

$$\frac{d}{dt}\gamma(t) = \frac{1}{(1+t)^\delta} h(\gamma(t), \zeta(t))$$

The output of this algorithm is then averaged:

$$\frac{d}{dt}\vartheta(t) = \frac{1}{1+t}\big(-\vartheta(t) + \gamma(t)\big)$$

Linear convergence under stability alone

1. There is a finite constant $\bar{\sigma}$ satisfying, $\boxed{\limsup_{t\to\infty} t\|\vartheta(t) - \vartheta^*\| \le \bar{\sigma}}$

# Polyak and Juditsky

Polyak and Juditsky[4, 2] obtain *optimal variance* for SA using a very simple approach: *High gain, and averaging*.

Deterministic P&J: Choose *high-gain*, $a(t) = 1/(1+t)^\delta$, with $\delta \in (0,1)$.

$$\frac{d}{dt}\gamma(t) = \frac{1}{(1+t)^\delta} h(\gamma(t), \zeta(t))$$

The output of this algorithm is then averaged:

$$\frac{d}{dt}\vartheta(t) = \frac{1}{1+t}\big(-\vartheta(t) + \gamma(t)\big)$$

Linear convergence under stability alone

1. There is a finite constant $\bar{\sigma}$ satisfying, $\boxed{\limsup_{t\to\infty} t\|\vartheta(t) - \vartheta^*\| \le \bar{\sigma}}$
2. Question: *Is $\bar{\sigma}$ minimal?*

# Conclusions

*Summary*:

- QSA is the most natural approach to approximate dynamic programming for deterministic systems, such as Q-learning[3].
- Stability is established using standard techniques
- Linear convergence is obtained under mild stability assumptions.

# Conclusions

*Summary*:

- QSA is the most natural approach to approximate dynamic programming for deterministic systems, such as Q-learning[3].
- Stability is established using standard techniques
- Linear convergence is obtained under mild stability assumptions.

*Current research*:

1. Open question: Can we extend the approach of Polyak and Juditsky to obtain optimal convergence?

# Conclusions

*Summary*:

- QSA is the most natural approach to approximate dynamic programming for deterministic systems, such as Q-learning[3].
- Stability is established using standard techniques
- Linear convergence is obtained under mild stability assumptions.

*Current research*:

1. Open question: Can we extend the approach of Polyak and Juditsky to obtain optimal convergence?
1. First we must answer, what is the optimal value of $\bar{\sigma}$?

# Conclusions

*Summary*:

- QSA is the most natural approach to approximate dynamic programming for deterministic systems, such as Q-learning[3].
- Stability is established using standard techniques
- Linear convergence is obtained under mild stability assumptions.

*Current research*:

1. Open question: Can we extend the approach of Polyak and Juditsky to obtain optimal convergence?
1. First we must answer, what is the optimal value of $\bar{\sigma}$?
2. Concentration on applications:
   - Further development of Q-learning
   - Applications to nonlinear filtering.

# References



D.P. Bertsekas and J. N. Tsitsiklis.
*Neuro-Dynamic Programming*.
Atena Scientific, Cambridge, Mass, 1996.


V. S. Borkar.
*Stochastic Approximation: A Dynamical Systems Viewpoint*.
Hindustan Book Agency and Cambridge University Press (jointly), Delhi, India and Cambridge, UK, 2008.


P. G. Mehta and S. P. Meyn.
Q-learning and Pontryagin's minimum principle.
In *Proc. of the 48th IEEE Conf. on Dec. and Control; held jointly with the 2009 28th Chinese Control Conference*, pages 3598–3605, Dec. 2009.


B. T. Polyak and A. B. Juditsky.
Acceleration of stochastic approximation by averaging.
*SIAM J. Control Optim.*, 30(4):838–855, 1992.


H. Robbins and S. Monro.
A stochastic approximation method.
*Annals of Mathematical Statistics*, 22:400–407, 1951.


C. J. C. H. Watkins and P. Dayan.
$Q$-learning.
*Machine Learning*, 8(3-4):279–292, 1992.