Continuous-time learning and optimization

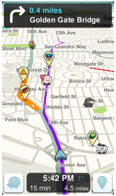
Walid Krichene

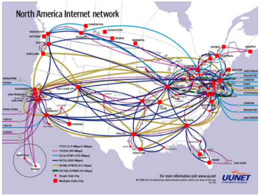
Electrical Engineering and Computer Sciences, UC Berkeley

June 9, 2016

Introduction

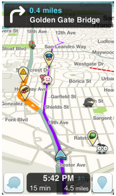
Coupled sequential decision problems: ubiquitous in Cyber-Physical Systems (CPS): Routing (transportation, communication), power networks.

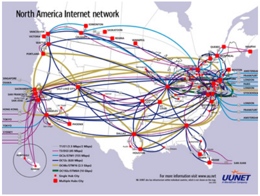




Introduction

Coupled sequential decision problems: ubiquitous in Cyber-Physical Systems (CPS): Routing (transportation, communication), power networks.





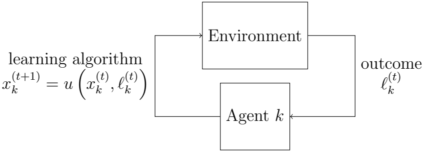
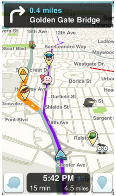
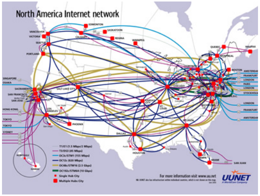


Figure: Sequential decision problem.

Introduction

Coupled sequential decision problems: ubiquitous in Cyber-Physical Systems (CPS): Routing (transportation, communication), power networks.





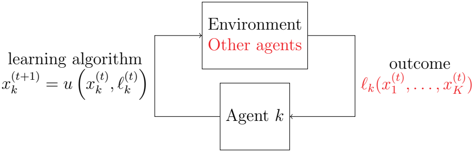


Figure: Coupled sequential decision problems.

Continuous-time dynamics

Design algorithms for learning and optimization in continuous-time.

Benefits

- Simple analysis.
- Provides insight into the discrete process.
- Streamlines design of new methods.
- Dynamics inspired from nature.

Accelerated Optimization in Continuous-Time

Constrained convex optimization

$$\begin{array}{lll} \text{minimize} & f(x) & (\text{convex } \nabla f \text{ Lipschitz}) \\ \text{subject to} & x \in \mathcal{X} & (\text{convex closed}) \end{array}$$

Gradient / Mirror descent [4]	$\mathcal{O}(1/k)$
Nesterov's accelerated method [5]	$\mathcal{O}(1/k^2)$

[4]A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience series in discrete mathematics. Wiley, 1983

[5]Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). *Soviet Mathematics Doklady*, 27(2):372–376, 1983

Accelerated Optimization in Continuous-Time

Constrained convex optimization

$$\begin{array}{lll} \text{minimize} & f(x) & (\text{convex } \nabla f \text{ Lipschitz}) \\ \text{subject to} & x \in \mathcal{X} & (\text{convex closed}) \end{array}$$

Gradient / Mirror descent [4]	$\mathcal{O}(1/k)$
Nesterov's accelerated method [5]	$\mathcal{O}(1/k^2)$

Dynamics	$\dot{X}(t) = -\nabla f(X(t))$	$x^{(k+1)} - x^{(k)} = -s\nabla f(x^{(k)})$
Lyapunov function	$E(t) := t(f(X(t)) - f^\star) + \frac{\|X(t)-x^\star\|^2}{2}$	$ks(f(x^{(k)}) - f^\star) + \frac{\|x^{(k)}-x^\star\|^2}{2}$
Convergence rate	$f(X(t)) - f^\star = \mathcal{O}(1/t)$	$f(x^{(k)}) - f^\star = \mathcal{O}(1/k)$

[4]A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience series in discrete mathematics. Wiley, 1983

[5]Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). *Soviet Mathematics Doklady*, 27(2):372–376, 1983

Lyapunov function

Candidate Lyapunov function parameterized by $r(t)$

$$L_r(t) = r(t)(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$$

$Z \in E^*$ dual variable, $z^\star$ its value at equilibrium, D_{ψ^*} Bregman divergence. If L_r is a Lyapunov function, then

$$f(X(t)) - f^\star \leq \frac{L_r(t)}{r(t)} \leq \frac{L_r(t_0)}{r(t)}$$

[2]W. Krichene, A. Bayen, and P. Bartlett. Accelerated mirror descent in continuous and discrete time. In *29th Annual Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015

Accelerated Mirror Descent

Accelerated Mirror Descent ODE

$$\text{AMD}_{w,\eta} \begin{cases} \dot{Z} = -\eta(t)\nabla f(X), \\ X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau} \\ \nabla\psi^*(z_0) = x_0. \end{cases}$$

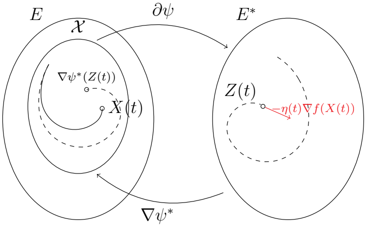


Figure: Z evolves in E^*, X is a weighted average of the mirrored trajectory $\nabla\psi^*(Z)$.

[3]W. Krichene, A. Bayen, and P. Bartlett. Adaptive averaging in accelerated descent dynamics. In *30th Annual Conference on Neural Information Processing Systems (NIPS), in review*, 2016

Accelerated Mirror Descent

Accelerated Mirror Descent ODE

$$\text{AMD}_{w,\eta} \begin{cases} \dot{Z} = -\eta(t)\nabla f(X), \\ X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau} \\ \nabla\psi^*(z_0) = x_0. \end{cases}$$

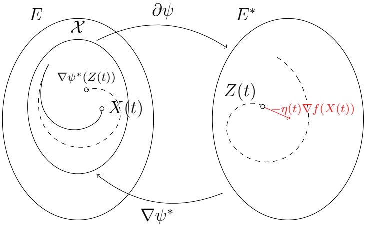


Figure: Z evolves in E^*, X is a weighted average of the mirrored trajectory $\nabla\psi^*(Z)$.

Convergence rate

If $\eta = \frac{wr}{W}$ and $w/W \geq r'/r$, then L_r is a Lyapunov function for $\text{AMD}_{w,\eta}$ and

$$f(X(t)) - f^\star \leq \frac{L_r(t_0)}{r(t)}$$

[3]W. Krichene, A. Bayen, and P. Bartlett. Adaptive averaging in accelerated descent dynamics. In *30th Annual Conference on Neural Information Processing Systems (NIPS), in review*, 2016

Example: accelerated replicator

Accelerated replicator ODE

$$\dot{\check{Z}}_i = \eta(t)\check{Z}_i \left(\langle \check{Z}, \nabla f(X)\rangle - \nabla_i f(X)\right) \qquad X = \frac{\int_0^t w(\tau)\check{Z}(\tau)d\tau}{\int_0^t w(\tau)d\tau}$$

Discretization

- t: continuous time
- k: discrete time
- s: step size
- Time correspondance: $t = k\sqrt{s}$.

Figure: Discretization

Heuristics to speed up convergence

- Restart when no progress [6]
- Adaptive averaging [3]

Figure: Restarting and adaptive averaging

[6]B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, pages 1–18, 2013

[3]W. Krichene, A. Bayen, and P. Bartlett. Adaptive averaging in accelerated descent dynamics. In *30th Annual Conference on Neural Information Processing Systems (NIPS), in review*, 2016

Optimization in nature

Continuous-time optimization can be found in nature.

- Physarum is shown to solve the shortest path problem (a linear program) [1].
- Dynamics modeled by biologists.
- Dynamics shown to be an instance of gradient descent applied to the shortest path problem (on some Riemannian manifold) [7].

Figure: Physarum can compute shortest paths [1]

---

[1]V. Bonifaci, K. Mehlhorn, and G. Varma. Physarum can compute shortest paths. *Journal of Theoretical Biology*, 309:121 – 133, 2012

[7]D. Straszak and N. K. Vishnoi. On a natural dynamics for linear programming. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16. ACM, 2016

References I

[1] V. Bonifaci, K. Mehlhorn, and G. Varma. Physarum can compute shortest paths. *Journal of Theoretical Biology*, 309:121 – 133, 2012.

[2] W. Krichene, A. Bayen, and P. Bartlett. Accelerated mirror descent in continuous and discrete time. In *29th Annual Conference on Neural Information Processing Systems (NIPS)*, Montreal, Canada, 2015.

[3] W. Krichene, A. Bayen, and P. Bartlett. Adaptive averaging in accelerated descent dynamics. In *30th Annual Conference on Neural Information Processing Systems (NIPS), in review*, 2016.

[4] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience series in discrete mathematics. Wiley, 1983.

[5] Y. Nesterov. A method of solving a convex programming problem with convergence rate o (1/k2). *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[6] B. O'Donoghue and E. Candès. Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, pages 1–18, 2013.

[7] D. Straszak and N. K. Vishnoi. On a natural dynamics for linear programming. In *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science*, ITCS '16. ACM, 2016.