Online Learning and Optimization
From Continuous to Discrete Time

Walid Krichene

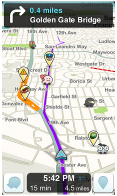
Electrical Engineering and Computer Sciences, UC Berkeley

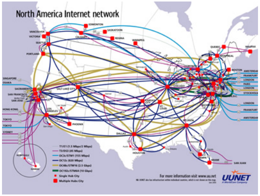
April 5, 2016

Introduction

Online Learning

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



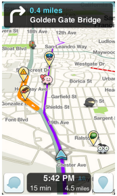


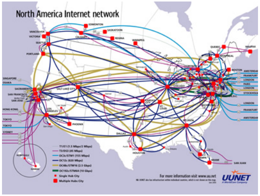
- Centralization impractical $\Rightarrow$ Distributed learning, e.g. learning in games.

Introduction

Online Learning

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





- Centralization impractical $\Rightarrow$ Distributed learning, e.g. learning in games.

Convex Optimization

- Data-driven decision problems.
- Size of data (dimension / sample size) makes higher-order methods prohibitively expensive.
- Active research on: {first-order, accelerated, stochastic} methods.

Introduction

Emerging idea

Design algorithms for online learning and optimization in continuous-time.

- Simple analysis.
- Provides insight into the discrete process.
- Streamlines design of new methods.

Continuous time $\leftrightarrow$ Discrete time

Outline

Outline

1. Discretizing the Replicator ODE

2. Accelerated Mirror Descent

Distributed learning in games

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $p \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+\mathbf{1})} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

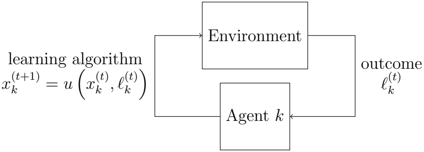


Figure: Sequential decision problem.

Distributed learning in games

Online Learning Model

for $t \in \mathbb{N}$ **do**
    Play $p \sim x_k^{(t)}$
    Discover $\ell_k^{(t)}$
    Update $x_k^{(t+\mathbf{1})} = u_k \left(x_k^{(t)}, \ell_k^{(t)} \right)$
end for

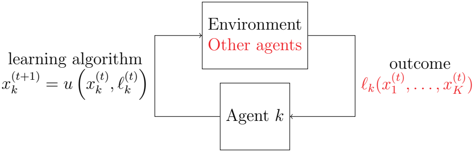


Figure: Coupled sequential decision problems.

Distributed learning in games

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $p \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+1)} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

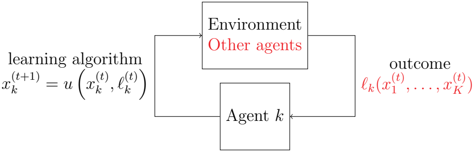


Figure: Coupled sequential decision problems.

- Equilibria: good description of system efficiency at steady-sate.
- Systems rarely operate at equilibrium.
- Study learning dynamics as
  1. A prescriptive model: How do we drive system to eq.
  2. A descriptive model: How would players behave in the game.

Distributed learning in games

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $p \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+\mathbf{1})} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

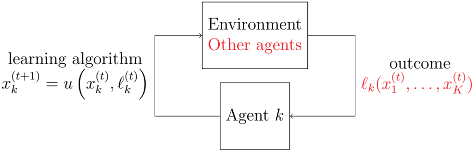


Figure: Coupled sequential decision problems.

- Equilibria: good description of system efficiency at steady-sate.
- Systems rarely operate at equilibrium.
- Study learning dynamics as
  1. A prescriptive model: How do we drive system to eq.
  2. A descriptive model: How would players behave in the game.

Goals

- Define classes of algorithms for which we can prove convergence.
- Robustness to stochastic perturbations.
- Heterogeneous learning (different agents use different algorithms).
- Convergence rates.

A brief review

Discrete time:

- Hannan consistency: [4]
- Hedge algorithm for two-player games: [3]
- Regret based algorithms: [5]
- Online learning in games: [2]

Continuous time:

- Evolution in populations: [13]
- Replicator dynamics in evolutionary game theory [15]
- No-regret dynamics for two player games [5]

---

[4]J. Hannan. Approximation to Bayes risk in repeated plays.
Contributions to the Theory of Games, 3:97–139, 1957

[3]Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights.
Games and Economic Behavior, 29(1):79–103, 1999

[5]S. Hart and A. Mas-Colell. A general class of adaptive strategies.
Journal of Economic Theory, 98(1):26 – 54, 2001

[2]N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games.*
Cambridge University Press, 2006

[13]W. H. Sandholm. *Population games and evolutionary dynamics.*
Economic learning and social evolution. Cambridge, Mass. MIT Press, 2010

[15]J. W. Weibull. *Evolutionary game theory.*
MIT press, 1997

[5]S. Hart and A. Mas-Colell. A general class of adaptive strategies.
Journal of Economic Theory, 98(1):26 – 54, 2001

Example: routing game

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $a \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+1)} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

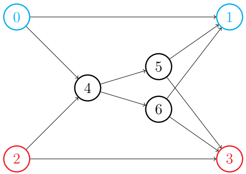


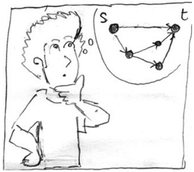
Figure: Routing game

Example: routing game

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $a \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+1)} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

$x_{\mathbf{1}}^{(t)} \in \Delta^{\mathbf{1}}$



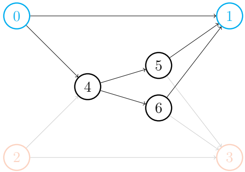


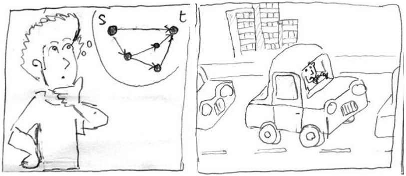
Figure: Routing game

Example: routing game

Online Learning Model

1: **for** $t \in \mathbb{N}$ **do**
2: Play $a \sim x_k^{(t)}$
3: Discover $\ell_k^{(t)}$
4: Update $x_k^{(t+1)} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
5: **end for**

$x_{\mathbf{1}}^{(t)} \in \Delta^{\mathbf{1}}$ Sample $a \sim x_{\mathbf{1}}^{(t)}$



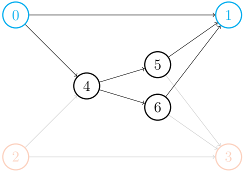


Figure: Routing game

Example: routing game

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $a \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+1)} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

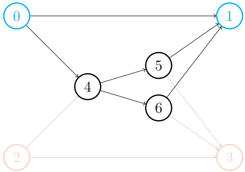
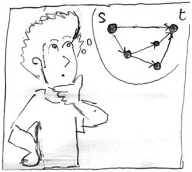
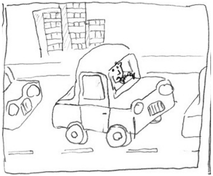


Figure: Routing game

$x_1^{(t)} \in \Delta^1$



Sample $a \sim x_1^{(t)}$



Discover $\ell_1^{(t)}$

Example: routing game

Online Learning Model

for $t \in \mathbb{N}$ **do**
  Play $a \sim x_k^{(t)}$
  Discover $\ell_k^{(t)}$
  Update $x_k^{(t+\mathbf{1})} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

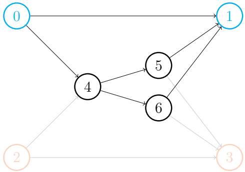
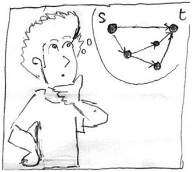
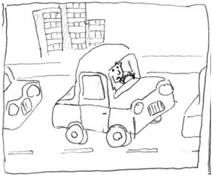


Figure: Routing game

$x_{\mathbf{1}}^{(t)} \in \Delta^{\mathbf{1}}$

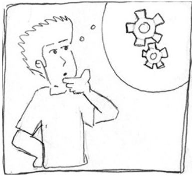


Sample $a \sim x_{\mathbf{1}}^{(t)}$



Discover $\ell_{\mathbf{1}}^{(t)}$

Update $x_{\mathbf{1}}^{(t+\mathbf{1})}$



Example: routing game

Online Learning Model

for $t \in \mathbb{N}$ **do**
    Play $a \sim x_k^{(t)}$
    Discover $\ell_k^{(t)}$
    Update $x_k^{(t+1)} = u_k\left(x_k^{(t)}, \ell_k^{(t)}\right)$
end for

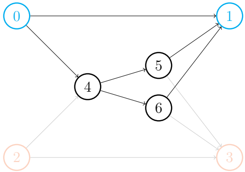
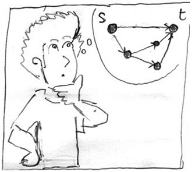
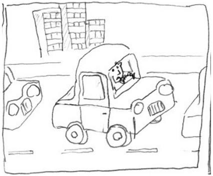


Figure: Routing game

$x_{\mathbf{1}}^{(t)} \in \Delta^{\mathbf{1}}$

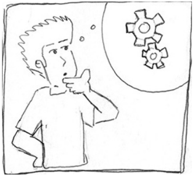


Sample $a \sim x_{\mathbf{1}}^{(t)}$



Discover $\ell_{\mathbf{1}}^{(t)}$

Update $x_{\mathbf{1}}^{(t+\mathbf{1})}$



Main problem

Define class of algorithms $\mathcal{C}$ such that

$$u_k \in \mathcal{C}\ \forall k \Rightarrow x^{(t)} \to \mathcal{X}^\star$$

Equilibria of the routing game

Write
$x = (x_{\mathcal{A}_1}, \dots, x_{\mathcal{A}_K}) \in \Delta^{\mathcal{A}_1} \times \cdots \times \Delta^{\mathcal{A}_K}$
$\ell(x) = (\ell_{\mathcal{A}_1}(x), \dots, \ell_{\mathcal{A}_K}(x))$

Nash equilibria $\mathcal{X}^\star$

$x^\star$ is a Nash equilibrium if for all k, paths in the support of $x^\star_{\mathcal{A}_k}$ have minimal loss.

$$\forall x,\ \langle \ell(x^\star), x - x^\star \rangle \geq 0$$

Equilibria of the routing game

Write
$x = (x_{\mathcal{A}_1}, \dots, x_{\mathcal{A}_K}) \in \Delta^{\mathcal{A}_1} \times \dots \times \Delta^{\mathcal{A}_K}$
$\ell(x) = (\ell_{\mathcal{A}_1}(x), \dots, \ell_{\mathcal{A}_K}(x))$

Nash equilibria $\mathcal{X}^\star$

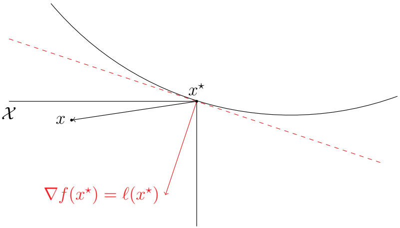
$x^\star$ is a Nash equilibrium if for all k, paths in the support of $x^\star_{\mathcal{A}_k}$ have minimal loss.

$$\forall x, \ \langle \ell(x^\star), x - x^\star \rangle \geq 0$$

Rosenthal potential

$\exists f$ convex such that $\nabla f(x) = \ell(x)$.

Nash condition	$\Leftrightarrow$	first order optimality
$\forall x, \ \langle \ell(x^\star), x - x^\star \rangle \geq 0$		$\forall x, \ \langle \nabla f(x^\star), x - x^\star \rangle \geq 0$



Stochastic approximation

Idea:

- View the learning dynamics as a discretization of an ODE.
- Study convergence of ODE.
- Relate convergence of discrete algorithm to convergence of ODE.

[15] J. W. Weibull. *Evolutionary game theory.*
MIT press, 1997

Stochastic approximation

Idea:

- View the learning dynamics as a discretization of an ODE.
- Study convergence of ODE.
- Relate convergence of discrete algorithm to convergence of ODE.

In Hedge $x_a^{(t+1)} \propto x_a^{(t)} e^{-\eta_t \ell_a^{(t)}}$, take $\eta_t \to 0$.

Replicator equation [15]

$$\forall a \in \mathcal{A}_k, \frac{dx_a}{dt} = x_a \left(\langle \ell(x), x \rangle - \ell_a(x) \right)$$

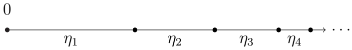


Figure: Underlying continuous time

[15] J. W. Weibull. *Evolutionary game theory.* MIT press, 1997

AREP dynamics: Approximate REPlicator

$$\frac{dx_a}{dt} = x_a \left(\langle \ell(x), x \rangle - \ell_a(x) \right)$$

Discretization of the continuous-time replicator dynamics

$$\frac{x_a^{(t+1)} - x_a^{(t)}}{\eta_t} = x_a^{(t)} \left(\left\langle \ell(x^{(t)}), x^{(t)} \right\rangle - \ell_a(x^{(t)}) \right) + U_a^{(t+1)}$$

[1] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de probabilités XXXIII*, pages 1–68. Springer, 1999

AREP dynamics: Approximate REPlicator

$$\frac{dx_a}{dt} = x_a\left(\langle \ell(x), x\rangle - \ell_a(x)\right)$$

Discretization of the continuous-time replicator dynamics

$$\frac{x_a^{(t+1)} - x_a^{(t)}}{\eta_t} = x_a^{(t)}\left(\left\langle \ell(x^{(t)}), x^{(t)}\right\rangle - \ell_a(x^{(t)})\right) + U_a^{(t+1)}$$

- η_t discretization time steps.

[1] M. Benaïm. Dynamics of stochastic approximation algorithms.
In *Séminaire de probabilités XXXIII*, pages 1–68. Springer, 1999

AREP dynamics: Approximate REPlicator

$$\frac{dx_a}{dt} = x_a \left(\langle \ell(x), x \rangle - \ell_a(x) \right)$$

Discretization of the continuous-time replicator dynamics

$$\frac{x_a^{(t+1)} - x_a^{(t)}}{\eta_t} = x_a^{(t)} \left(\left\langle \ell(x^{(t)}), x^{(t)} \right\rangle - \ell_a(x^{(t)}) \right) + U_a^{(t+1)}$$

- η_t discretization time steps.
- $(U^{(t)})_{t \geq 1}$ perturbations that satisfy for all $T > 0$, $\lim_{\tau_1 \to \infty} \max_{\tau_2 : \sum_{t=\tau_1}^{\tau_2} \eta_t < T} \left\| \sum_{t=\tau_1}^{\tau_2} \eta_t U^{(t+1)} \right\| = 0$

(a sufficient condition is that $\exists q \geq 2$: $\sup_\tau \mathbb{E} \| U^{(\tau)} \|^q < \infty$ and $\sum_\tau \eta_\tau^{1+\frac{q}{2}} < \infty$)

[1] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de probabilités XXXIII*, pages 1–68. Springer, 1999

Convergence to Nash equilibria

Theorem [6]

In convex potential games, under AREP updates, if $\eta_t \downarrow 0$ and $\sum \eta_t = \infty$, then

$$x^{(t)} \to \mathcal{X}^\star \text{ a.s.}$$

[6] W. Krichene, B. Drighès, and A. Bayen. Learning nash equilibria in congestion games. *SIAM Journal on Control and Optimization (SICON), to appear*, 2014

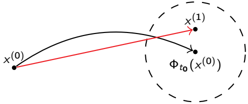
Convergence to Nash equilibria

Theorem [6]

In convex potential games, under AREP updates, if $\eta_t \downarrow 0$ and $\sum \eta_t = \infty$, then

$$x^{(t)} \to \mathcal{X}^\star \text{ a.s.}$$

- Affine interpolation of $x^{(t)}$ is an asymptotic pseudo trajectory of ODE.



- Use f as a Lyapunov function. proof details

[6] W. Krichene, B. Drighès, and A. Bayen. Learning nash equilibria in congestion games. *SIAM Journal on Control and Optimization (SICON), to appear*, 2014

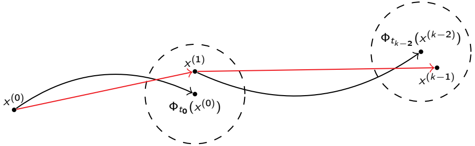
Convergence to Nash equilibria

Theorem [6]

In convex potential games, under AREP updates, if $\eta_t \downarrow 0$ and $\sum \eta_t = \infty$, then

$$x^{(t)} \to \mathcal{X}^\star \text{ a.s.}$$

- Affine interpolation of $x^{(t)}$ is an asymptotic pseudo trajectory of ODE.



- Use f as a Lyapunov function. proof details

However, No convergence rates.

[6] W. Krichene, B. Drighès, and A. Bayen. Learning nash equilibria in congestion games. *SIAM Journal on Control and Optimization (SICON), to appear*, 2014

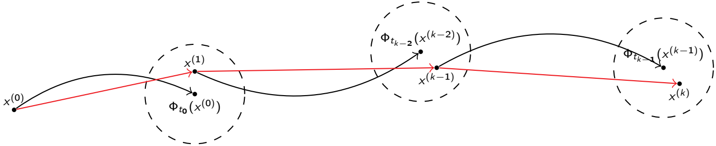
Convergence to Nash equilibria

Theorem [6]

In convex potential games, under AREP updates, if $\eta_t \downarrow 0$ and $\sum \eta_t = \infty$, then

$$x^{(t)} \to \mathcal{X}^\star \text{ a.s.}$$

- Affine interpolation of $x^{(t)}$ is an asymptotic pseudo trajectory of ODE.



- Use f as a Lyapunov function. proof details

However, No convergence rates.

[6] W. Krichene, B. Drighès, and A. Bayen. Learning nash equilibria in congestion games. *SIAM Journal on Control and Optimization (SICON), to appear*, 2014

Asymptotic Pseudo Trajectory

Figure: Discrete (Hedge) and continuous (Replicator) trajectories

Numerical example

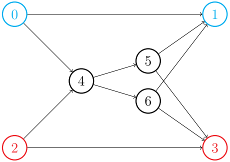
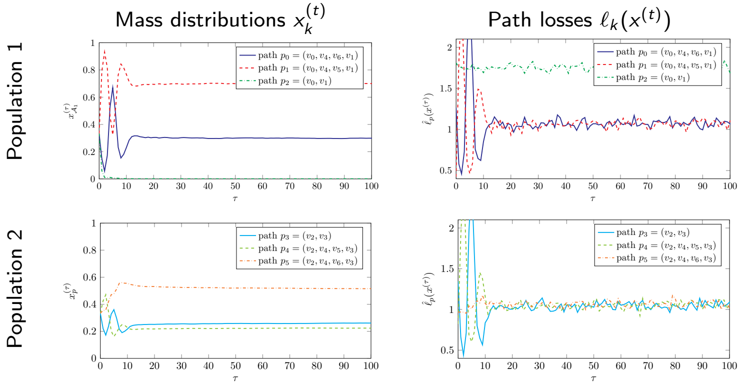
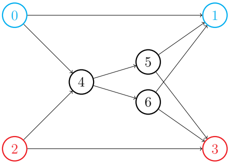


Figure: Example with strongly convex potential.

- Centered Gaussian noise on edges.
- Population 1: Hedge with $\eta_t^1 = t^{-1}$
- Population 2: Hedge with $\eta_t^2 = t^{-1}$



Numerical example



- Centered Gaussian noise on edges.
- Population 1: Hedge with $\eta_t^1 = t^{-1}$
- Population 2: Hedge with $\eta_t^2 = t^{-1}$

Figure: Example with strongly convex potential.

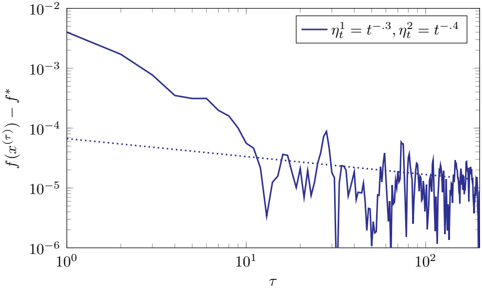
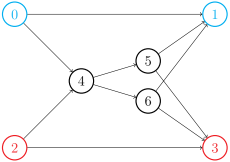


Figure: Potential values.

For $\eta_t^k = \frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0,1)$, $\mathbb{E}\left[f(x^{(t)})\right] - f^\star = O\left(\sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}}\right)$

Numerical example



- Centered Gaussian noise on edges.
- Population 1: Hedge with $\eta_t^1 = t^{-1}$
- Population 2: Hedge with $\eta_t^2 = t^{-1}$

Figure: Example with strongly convex potential.

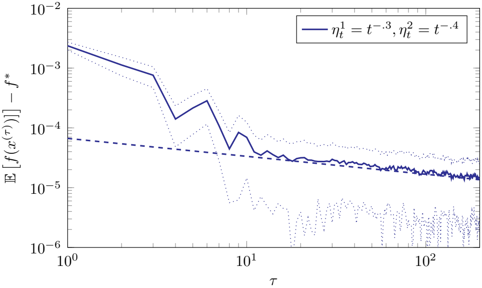


Figure: Potential values.

For $\eta_t^k = \frac{\theta_k}{t^{\alpha_k}}$, $\alpha_k \in (0,1)$, $\mathbb{E}\left[f(x^{(t)})\right] - f^\star = O\left(\sum_k \frac{\log t}{t^{\min(\alpha_k, 1-\alpha_k)}}\right)$

Outline

1. Discretizing the Replicator ODE
2. Accelerated Mirror Descent

First order optimization: from continuous to discrete time

Constrained convex optimization

$$\begin{aligned} &\text{minimize} && f(x) \\ &\text{subject to} && x \in \mathcal{X} \end{aligned}$$

- f is convex differentiable, L_f smooth (i.e. ∇f is L_f Lipschitz).
- $\mathcal{X}$ is convex closed.
- First-order: can evaluate $f(x)$ and $\nabla f(x)$.

Gradient descent	$\mathcal{O}(1/k)$
Mirror descent [9] Dual Averaging [11]	$\mathcal{O}(1/k)$
Nesterov's accelerated method [10]	$\mathcal{O}(1/k^2)$

Goal: unified approach to derive these algorithms.

- Design ODE in continuous time using Lyapunov argument.
- Discretize.

---

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

[11]Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009

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

From Gradient Descent to Mirror Descent

Gradient descent is discretization of

Gradient descent ODE

$$\dot{X} = -\nabla f(X)$$

Converges in $\mathcal{O}(1/t)$.

Proof idea: define $D(X(t), x^\star) = \frac{1}{2}\|X(t) - x^\star\|^2$.

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

From Gradient Descent to Mirror Descent

Gradient descent is discretization of

Gradient descent ODE

$$\dot{X} = -\nabla f(X)$$

Converges in $\mathcal{O}(1/t)$.

Proof idea: define $D(X(t), x^\star) = \frac{1}{2}\|X(t) - x^\star\|^2$.
Nemirovski and Yudin [9]

1. Start from function on the dual space

$$D_{\psi^*}(Z, z^\star) = \psi^*(Z) - \psi^*(z^\star) - \langle \nabla\psi^*(z^\star), Z - z^\star \rangle$$

2. Design dynamics to make it a Lyapunov function.

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

From Gradient Descent to Mirror Descent

Mirror descent ODE

$$\begin{cases} \dot{Z} = -\nabla f(X) \\ X = \nabla \psi^*(Z) \end{cases}$$

Converges in $\mathcal{O}(1/t)$.

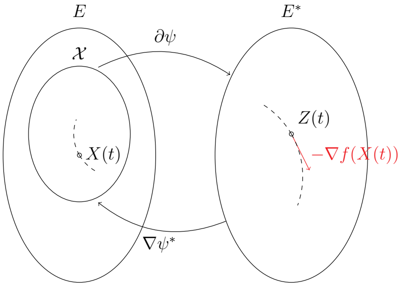


Figure: Illustration of Mirror Descent

ψ^* is defined and differentiable on E^*, $\nabla\psi^*$ maps E^* to $\mathcal{X}$. More on $\nabla\psi^*$

An ODE interpretation of Nesterov's method

Su et al. [14]: for unconstrained problems

1. Nesterov's method is discretization of

$$\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0$$

[14]W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

An ODE interpretation of Nesterov's method

Su et al. [14]: for unconstrained problems

1. Nesterov's method is discretization of

$$\ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0$$

2. Proved convergence at $\mathcal{O}(1/t^2)$ rate. Argument: Lyapunov function

$$\frac{t^2}{r}(f(X) - f^\star) + \frac{r}{2}\|X + \frac{t}{r}\dot{X} - x^\star\|_2^2$$

[14]W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

Accelerated Mirror Descent in continuous time

We start from a Lyapunov function [7]

$$V(X, Z, t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$$

$r \geq 2$, a parameter, $Z \in E^*$, $z^\star$ its value at equilibrium.

[7]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 in continuous time

We start from a Lyapunov function [7]

$$V(X, Z, t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$$

$r \geq 2$, a parameter, $Z \in E^*$, $z^\star$ its value at equilibrium.

AMD ODE

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla \psi^*(Z) - X), \end{cases} \tag{1}$$

If (X, Z) is a solution to ODE (1), then V is a Lyapunov function.

[7]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 in continuous time

We start from a Lyapunov function [7]

$$V(X, Z, t) = \frac{t^2}{r^2}(f(X(t)) - f^\star) + D_{\psi^*}(Z(t), z^\star)$$

$r \geq 2$, a parameter, $Z \in E^*$, $z^\star$ its value at equilibrium.

AMD ODE

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases} \tag{1}$$

If (X, Z) is a solution to ODE (1), then V is a Lyapunov function.

Consequence: convergence rate

$$f(X(t)) - f^\star \leq \frac{r^2 D_{\psi^*}(z_0, z^\star)}{t^2}$$

Proof: $f(X(t)) - f^\star \leq \frac{r^2 V(X(t),Z(t),t)}{t^2} \leq \frac{rV(x_0,z_0,0)}{t^2} = \frac{r^2 D_{\psi^*}(z_0,z^\star)}{t^2}$

---

[7]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

Averaging Interpretation

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases}$$

Averaging interpretation

Second equation equivalent to

$$X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}$$

with $w(\tau) = \tau^{r-1}$.

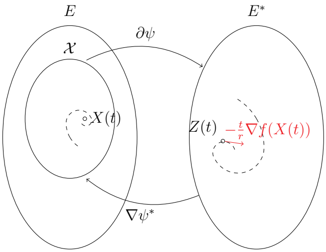


Figure: Averaging interpretation: Z evolves in E^*, X is a weighted average of the mirrored trajectory $\nabla\psi^*(Z)$.

[8]W. Krichene, A. Bayen, and P. Bartlett. A Lyapunov approach to first-order methods for convex optimization, in continuous and discrete time.
SIAM Journal on Optimization (SIOPT), submitted, December 2015

Averaging Interpretation

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases}$$

Averaging interpretation

Second equation equivalent to

$$X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}$$

with $w(\tau) = \tau^{r-1}$.

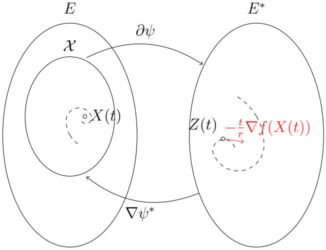


Figure: Averaging interpretation: Z evolves in E^*, X is a weighted average of the mirrored trajectory $\nabla\psi^*(Z)$.

General averaging[8]

If $W(t) = \int_0^t w(\tau)d\tau$, and $\frac{w}{W} \geq \frac{2}{t}$, then V is Lyapunov under

$$\dot{Z} = -\frac{w}{W}\frac{t^2}{r^2}\nabla f(X)$$

[8]W. Krichene, A. Bayen, and P. Bartlett. A Lyapunov approach to first-order methods for convex optimization, in continuous and discrete time. *SIAM Journal on Optimization (SIOPT), submitted*, December 2015

Example: accelerated entropic descent on the simplex

Suppose the feasible set is $\mathcal{X} = \Delta^n = \{x \in \mathbb{R}^n_+ : \sum_i x_i = 1\}$.

$$\psi(x) = \sum_i x_i \ln x_i + \delta(x|\Delta), \qquad \psi^*(z) = \ln \sum_i e^{z_i}, \qquad \nabla\psi^*(z)_i = \frac{e^{z_i}}{\sum_i e^{z_i}},$$

Accelerated replicator ODE

$$\begin{cases} \dot{\tilde{Z}}_i = \tilde{Z}_i \left(\left\langle \tilde{Z}, \nabla f(X) \right\rangle - \nabla_i f(X) \right) \\ X = \frac{\int_0^t \tau^{r-1} \tilde{Z}(\tau) d\tau}{\int_0^t \tau^{r-1} d\tau} \end{cases}$$

Numerical Example

Figure: Accelerated entropic descent on a quadratic on the simplex.

Damped oscillator interpretation

Damped nonlinear oscillator

Accelerated mirror descent ODE is equivalent to

$$\ddot{X} + \frac{r+1}{t}\dot{X} = -\nabla^2\psi^*(Z)\nabla f(X)$$

Damped oscillator interpretation

Damped nonlinear oscillator

Accelerated mirror descent ODE is equivalent to

$$\ddot{X} + \frac{r+1}{t}\dot{X} = -\nabla^2\psi^*(Z)\nabla f(X)$$

- Special case: $\ddot{X} + \frac{r+1}{t}\dot{X} = -\nabla f(X)$
- $\frac{r+1}{t}\dot{X}$: vanishing friction term.

Effect of the parameter r

$$\ddot{X} + \frac{r+1}{t}\dot{X} = -\nabla^2\psi^*(Z)\nabla f(X)$$

Figure: Effect of the parameter $r \in [2, 50]$.

Effect of $\nabla^2\psi^*(Z)$

$$\ddot{X} + \frac{r+1}{t}\dot{X} = -\nabla^2\psi^*(Z)\nabla f(X)$$

Figure: Flow field $x \mapsto \nabla^{\mathbf{2}}\psi^*(Z(t))\nabla f(x)$, along the solution trajectory Z

Existence and uniqueness of the solution

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases}$$

Solution

Suppose ∇f and $\nabla\psi^*$ are Lipschitz. Then ODE system (1) has a unique solution defined on $[0, +\infty)$, and the solution remains in $\mathcal{X}$.

Existence and uniqueness of the solution

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases}$$

Solution

Suppose ∇f and $\nabla\psi^*$ are Lipschitz. Then ODE system (1) has a unique solution defined on $[0, +\infty)$, and the solution remains in $\mathcal{X}$.

Proof sketch: Would like to invoke Cauchy-Lipschitz theorem (Picard-Lindelöf), but singularity at 0.

Existence and uniqueness of the solution

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases}$$

Solution

Suppose ∇f and $\nabla\psi^*$ are Lipschitz. Then ODE system (1) has a unique solution defined on $[0, +\infty)$, and the solution remains in $\mathcal{X}$.

Proof sketch: Would like to invoke Cauchy-Lipschitz theorem (Picard-Lindelöf), but singularity at 0.

1. Define family of “smoothed” ODEs:

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X), \end{cases}$$

Existence and uniqueness of the solution

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases}$$

Solution

Suppose ∇f and $\nabla\psi^*$ are Lipschitz. Then ODE system (1) has a unique solution defined on $[0, +\infty)$, and the solution remains in $\mathcal{X}$.

Proof sketch: Would like to invoke Cauchy-Lipschitz theorem (Picard-Lindelöf), but singularity at 0.

1. Define family of "smoothed" ODEs:

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X), \end{cases}$$

2. Extract a converging subsequence. Its limit is a solution to (1).

Discretization

Time correspondence: $t = k\sqrt{s}$, for a step size s. First attempt:

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X), \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \end{cases} \qquad \begin{cases} \frac{z^{(k+\mathbf{1})} - z^{(k)}}{\sqrt{s}} = -\frac{k\sqrt{s}}{r}\nabla f(x^{(k)}) \\ \frac{x^{(k+\mathbf{1})} - x^{(k)}}{\sqrt{s}} = \frac{r}{k\sqrt{s}}\left(\nabla\psi^*(z^{(k+1)}) - x^{(k+1)}\right). \end{cases}$$

Candidate Lyapunov function:

$$E^{(k)} = V(x^{(k)}, z^{(k)}, k\sqrt{s}).$$

Discrete AMD algorithm.

Accelerated mirror descent with distance generating function ψ^*, regularizer R

1: Initialize $\tilde{x}^{(0)} = x_0$, $\tilde{z}^{(0)} = x_0$
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad \tilde{z}^{(k+1)} = \arg\min_{\tilde{z} \in \mathcal{X}} \frac{kr}{s} \left\langle \nabla f(x^{(k)}), \tilde{z} \right\rangle + D_\psi(\tilde{z}, x^{(k)})$
4: $\quad \tilde{x}^{(k+1)} = \arg\min_{\tilde{x} \in \mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(k)})$
5: $\quad x^{(k+1)} = \lambda_k \tilde{z}^{(k+1)} + (1 - \lambda_k)\tilde{x}^{(k+1)}$, with $\lambda_k = \frac{r}{r+k}$.
6: **end for**

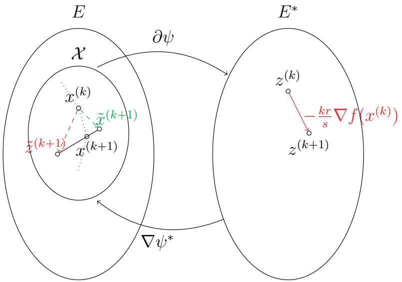
- R regularizer function, assumed strongly convex and smooth.

Discrete AMD algorithm.

Accelerated mirror descent with distance generating function ψ^*, regularizer R

1: Initialize $\tilde{x}^{(0)} = x_0$, $\tilde{z}^{(0)} = x_0$
2: **for** $k \in \mathbb{N}$ **do**
3: $\quad \tilde{z}^{(k+1)} = \arg\min_{\tilde{z} \in \mathcal{X}} \frac{kr}{s} \langle \nabla f(x^{(k)}), \tilde{z} \rangle + D_\psi(\tilde{z}, x^{(k)})$
4: $\quad \tilde{x}^{(k+1)} = \arg\min_{\tilde{x} \in \mathcal{X}} \gamma s \langle \nabla f(x^{(k)}), \tilde{x} \rangle + R(\tilde{x}, x^{(k)})$
5: $\quad x^{(k+1)} = \lambda_k \tilde{z}^{(k+1)} + (1 - \lambda_k)\tilde{x}^{(k+1)}$, with $\lambda_k = \frac{r}{r+k}$.
6: **end for**

- R regularizer function, assumed strongly convex and smooth.
- Modified scheme is consistent with the ODE. Idea: $\tilde{x}^{(k)} = x^{(k)} + \mathcal{O}(s)$.



Convergence rate

Convergence rate

If $\gamma \geq L_f L_{\psi^*}$ and $s \leq \frac{\ell_R}{2L_f\gamma}$, then

$$f(\tilde{x}^{(k)}) - f^\star \leq C/k^2,$$

where $C = \frac{r^2 D_{\psi^*}(z_0, z^\star)}{s} + f(x_0) - f^\star$.

Proof: $\tilde{E}^{(k)} = V(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s})$ is a Lyapunov function.

Restarting

Restart the algorithm when a certain condition is met.

- Gradient restart: $\left\langle x^{(k+1)} - x^{(k)}, \nabla f(x^{(k)}) \right\rangle > 0$
- Speed restart: $\|x^{(k+1)} - x^{(k)}\| < \|x^{(k)} - x^{(k-1)}\|$

Algorithm 1 Accelerated mirror descent with restart

Initialize $l = 0$, $\tilde{x}^{(0)} = \tilde{z}^{(0)} = x_0$.
for $k \in \mathbb{N}$ **do**
    $\tilde{z}^{(k+1)} = \arg\min_{\tilde{z} \in \mathcal{X}} \frac{lr}{s} \left\langle \nabla f(x^{(k)}), \tilde{z} \right\rangle + D_\psi(\tilde{z}, x^{(k)})$
    $\tilde{x}^{(k+1)} = \arg\min_{\tilde{x} \in \mathcal{X}} \gamma s \left\langle \nabla f(x^{(k)}), \tilde{x} \right\rangle + R(\tilde{x}, x^{(k)})$
    $x^{(k+1)} = \lambda_l \tilde{z}^{(k+1)} + (1 - \lambda_l)\tilde{x}^{(k+1)}$, with $\lambda_l = \frac{r}{r+l}$.
    $l \leftarrow l + 1$
    if Restart condition **then**
        $\tilde{z}^{(k+1)} \leftarrow x^{(k+1)}$, $l \leftarrow 0$
    end if
end for

Illustration of restarting

Figure: Illustration of restarting

Example with a weakly convex function

Figure: Example with a weakly convex function. The black segment shows $\arg\min f$. Observe that each method converges to some point $x^\star \in \arg\min f$.

Dynamical systems approach to optimization

Paradigm

- Design ODE in continuous-time.
- Streamline the discretization.

For practitioners: Use off-the-shelf numerical methods to discretize the ODE.

Dynamical systems approach to optimization

Paradigm

- Design ODE in continuous-time.
- Streamline the discretization.

For practitioners: Use off-the-shelf numerical methods to discretize the ODE.

Develop the theory:

- Rigorous analysis of effect of r. Adaptive r?
- Study restarting heuristics.

Dynamical systems approach to optimization

Paradigm

- Design ODE in continuous-time.
- Streamline the discretization.

For practitioners: Use off-the-shelf numerical methods to discretize the ODE.

Develop the theory:

- Rigorous analysis of effect of r. Adaptive r?
- Study restarting heuristics.
- Monotone operators.
- Composite optimization

$$\min_{x \in \mathcal{X}} f(x) + g(x)$$

where ∇f is Lipschitz and g is a general convex function.

Thank you!

References I

[1] M. Benaïm. Dynamics of stochastic approximation algorithms. In *Séminaire de probabilités XXXIII*, pages 1–68. Springer, 1999.

[2] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

[3] Y. Freund and R. E. Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1):79–103, 1999.

[4] J. Hannan. Approximation to Bayes risk in repeated plays. *Contributions to the Theory of Games*, 3:97–139, 1957.

[5] S. Hart and A. Mas-Colell. A general class of adaptive strategies. *Journal of Economic Theory*, 98(1):26 – 54, 2001.

[6] W. Krichene, B. Drighès, and A. Bayen. Learning nash equilibria in congestion games. *SIAM Journal on Control and Optimization (SICON), to appear*, 2014.

[7] 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.

[8] W. Krichene, A. Bayen, and P. Bartlett. A Lyapunov approach to first-order methods for convex optimization, in continuous and discrete time. *SIAM Journal on Optimization (SIOPT), submitted*, December 2015.

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

References II

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

[11] Y. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, 120(1):221–259, 2009.

[12] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1997.

[13] W. H. Sandholm. *Population games and evolutionary dynamics*. Economic learning and social evolution. Cambridge, Mass. MIT Press, 2010.

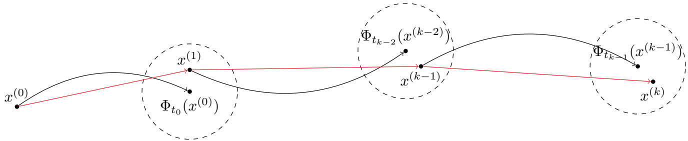
[14] W. Su, S. Boyd, and E. Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *NIPS*, 2014.

[15] J. W. Weibull. *Evolutionary game theory*. MIT press, 1997.

AREP convergence proof

Back

- Affine interpolation of $x^{(t)}$ is an asymptotic pseudo trajectory.



- The set of limit points of an APT is internally chain transitive ICT.
- If Γ is compact invariant, and has a Lyapunov function f with $\operatorname{int} f(\Gamma) = \emptyset$, then $\forall L$ ICT, Γ, and f is constant on L.
- In particular, f is constant on $L(x^{(t)})$, so $f(x^{(t)})$ converges.

More on the mirror operator $\nabla\psi^*$

▸ Back to mirror descent

Consider a pair of closed conjugate convex functions ψ, ψ^*

- $\psi : \mathcal{X} \to \mathbb{R}$

[12]R. Rockafellar. *Convex Analysis.*
Princeton University Press, 1997

More on the mirror operator $\nabla\psi^*$

Back to mirror descent

Consider a pair of closed conjugate convex functions ψ, ψ^*

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x\in\mathcal{X}} \langle z, x\rangle - \psi(x)$

[12]R. Rockafellar. *Convex Analysis.*
Princeton University Press, 1997

More on the mirror operator $\nabla\psi^*$

Back to mirror descent

Consider a pair of closed conjugate convex functions ψ, ψ^*

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
- $\partial\psi^*(z) = \arg\max_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  (so $\partial\psi^*$ naturally maps into $\mathcal{X}$).

[12]R. Rockafellar. *Convex Analysis.*
Princeton University Press, 1997

More on the mirror operator $\nabla\psi^*$

Back to mirror descent

Consider a pair of closed conjugate convex functions ψ, ψ^*

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x\in\mathcal{X}} \langle z, x\rangle - \psi(x)$
- $\partial\psi^*(z) = \arg\max_{x\in\mathcal{X}} \langle z, x\rangle - \psi(x)$
  (so $\partial\psi^*$ naturally maps into $\mathcal{X}$).

Mirror operator

If $\psi : \mathcal{X} \to \mathbb{R}$ is convex, closed, (essentially) strongly convex, such that epi f contains no non-vertical half-lines, then ψ^* is finite differentiable on E^* and $\nabla\psi^* : E^* \to \mathcal{X}$.

---

[12]R. Rockafellar. *Convex Analysis.*
Princeton University Press, 1997

The mirror operator $\nabla\psi^*$

Back to mirror descent

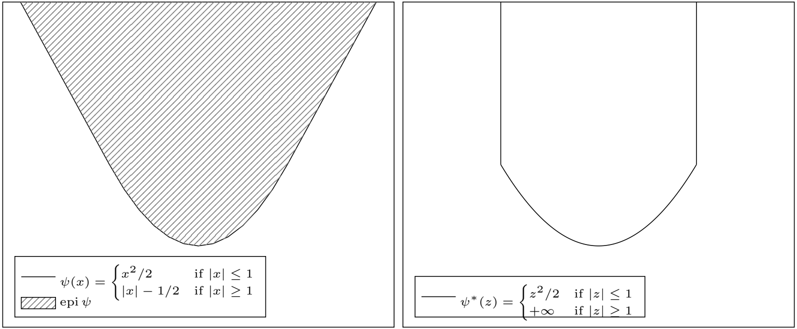


Figure: Example of dual distance generating functions ψ and ψ^*.

Application to load balancing

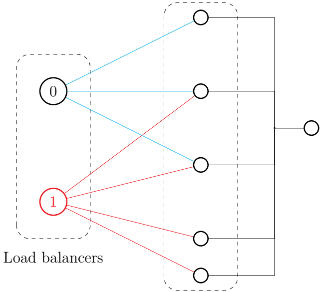


Figure: Load balancing problem.

- Modeled using a routing game.
- Can be solved using AMD.
- Acceleration leads to oscillation, undesirable.
- Use restarting heuristics to detect and alleviate oscillations.