A Lyapunov approach to Accelerated First Order Optimization In Continuous and Discrete Time

Walid Krichene Alexandre Bayen Peter Bartlett

Electrical Engineering and Computer Sciences, UC Berkeley

October 30, 2015

First order optimization

Consider the constrained optimization problem

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

- f is convex differentiable, L_f smooth (i.e. ∇f is L_f Lipschitz).
- $\mathcal{X}$ is convex closed.

First order optimization

Consider the constrained optimization problem

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

- 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 $\nabla f(x)$.

First order optimization

Consider the constrained optimization problem

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

- 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 $\nabla f(x)$.
(Evaluating / inverting Hessians is prohibitive in high dimensions. An alternative is to approximate the Hessian: LBFGS, random projections, etc.)

First order optimization

Gradient descent	$x^{(k+1)} = x^{(k)} - s\nabla f(x^{(k)})$	$\mathcal{O}(1/k)$
Mirror descent [2] Dual Averaging [4]	$\begin{cases} x^{(k+1)} = \nabla\psi^*(z^{(k)}) \\ z^{(k+1)} = z^{(k)} - s_k\nabla f(x^{(k)}) \end{cases}$	$\mathcal{O}(1/k)$
Nesterov's accelerated method [3]	$\begin{cases} y^{(k+1)} = x^{(k)} - s\nabla f(x^{(k)}) \\ x^{(k+1)} = y^{(k+1)} + \beta_k[y^{(k+1)} - y^{(k)}] \end{cases}$	$\mathcal{O}(1/k^2)$

Table: First-order methods

---

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

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

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

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

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

[1]Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming.

First order optimization

Gradient descent	$x^{(k+1)} = x^{(k)} - s\nabla f(x^{(k)})$	$\mathcal{O}(1/k)$
Mirror descent [2] Dual Averaging [4]	$\begin{cases} x^{(k+1)} = \nabla\psi^*(z^{(k)}) \\ z^{(k+1)} = z^{(k)} - s_k\nabla f(x^{(k)}) \end{cases}$	$\mathcal{O}(1/k)$
Nesterov's accelerated method [3]	$\begin{cases} y^{(k+1)} = x^{(k)} - s\nabla f(x^{(k)}) \\ x^{(k+1)} = y^{(k+1)} + \beta_k[y^{(k+1)} - y^{(k)}] \end{cases}$	$\mathcal{O}(1/k^2)$

Table: First-order methods

Other methods: Frank-Wolfe [1]

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

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

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

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

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

[1]Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming.

A continuous-time motivation

Motivation from continuous-time systems:

- Gradient descent: discretization of $\dot{X} = -\nabla f(X)$

---

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

[7]Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

A continuous-time motivation

Motivation from continuous-time systems:

- Gradient descent: discretization of $\dot{X} = -\nabla f(X)$
- Mirror descent [2].

---

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

[7]Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

A continuous-time motivation

Motivation from continuous-time systems:

- Gradient descent: discretization of $\dot{X} = -\nabla f(X)$
- Mirror descent [2].
- Nesterov's method: a recent continuous-time interpretation for the unconstrained Euclidean case [7]

---

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

[7]Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

A continuous-time motivation

Motivation from continuous-time systems:

- Gradient descent: discretization of $\dot{X} = -\nabla f(X)$
- Mirror descent [2].
- Nesterov's method: a recent continuous-time interpretation for the unconstrained Euclidean case [7]

We will unify accelerated methods in a family of ODEs.

- Continuous time: existence and uniqueness of solutions, convergence rates.
- Discrete time: convergence rates, heuristics for faster rates.

---

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

[7]Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

Introduction:
Mirror Descent and Nesterov

Gradient descent ODE

Gradient descent is discretization of

Gradient descent ODE

$$\begin{cases} \dot{X} = -\nabla f(X) \\ X(0) = x_0 \end{cases}$$

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

Gradient descent ODE

Gradient descent is discretization of

Gradient descent ODE

$$\begin{cases} \dot{X} = -\nabla f(X) \\ X(0) = x_0 \end{cases}$$

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

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

$$\begin{aligned} \frac{d}{dt} D(X(t), x^\star) &= \left\langle \dot{X}, X - x^\star \right\rangle \\ &= -\left\langle \nabla f(X), X - x^\star \right\rangle \\ &\leq -(f(X) - f^\star) \end{aligned}$$

Gradient descent ODE

Gradient descent is discretization of

Gradient descent ODE

$$\begin{cases} \dot{X} = -\nabla f(X) \\ X(0) = x_0 \end{cases}$$

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

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

$$\begin{aligned} \frac{d}{dt} D(X(t), x^\star) &= \left\langle \dot{X}, X - x^\star \right\rangle \\ &= -\left\langle \nabla f(X), X - x^\star \right\rangle \\ &\leq -(f(X) - f^\star) \end{aligned}$$

thus

$$\frac{1}{t}\int_0^t f(X(\tau))d\tau - f^\star \leq \frac{D(x_0, x^\star)}{t}$$

Mirror Descent ODE

Nemirovski and Yudin [2]

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.

---

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

Mirror Descent ODE

Nemirovski and Yudin [2]

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.

$$\frac{d}{dt} D_{\psi^*}(Z, z^\star) = \left\langle \nabla\psi^*(Z), \dot{Z} \right\rangle - \left\langle \nabla\psi^*(z^\star), \dot{Z} \right\rangle$$

---

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

Mirror Descent ODE

Nemirovski and Yudin [2]

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.

$$\frac{d}{dt} D_{\psi^*}(Z, z^\star) = \left\langle \nabla\psi^*(Z), \dot{Z} \right\rangle - \left\langle \nabla\psi^*(z^\star), \dot{Z} \right\rangle$$

Mirror descent ODE

$$\begin{cases} \dot{Z} = -\nabla f(X) \\ X = \nabla\psi^*(Z) \\ X(0) = x_0, \;\; Z(0) \in (\nabla\psi^*)^{-1}(x_0) \end{cases}$$

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

proof:

$$\frac{d}{dt} D_{\psi^*}(Z, z^\star) = \langle -\nabla f(X), X - x^\star \rangle$$

~~then same argument as gradient descent~~.

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

The mirror operator $\nabla\psi^*$

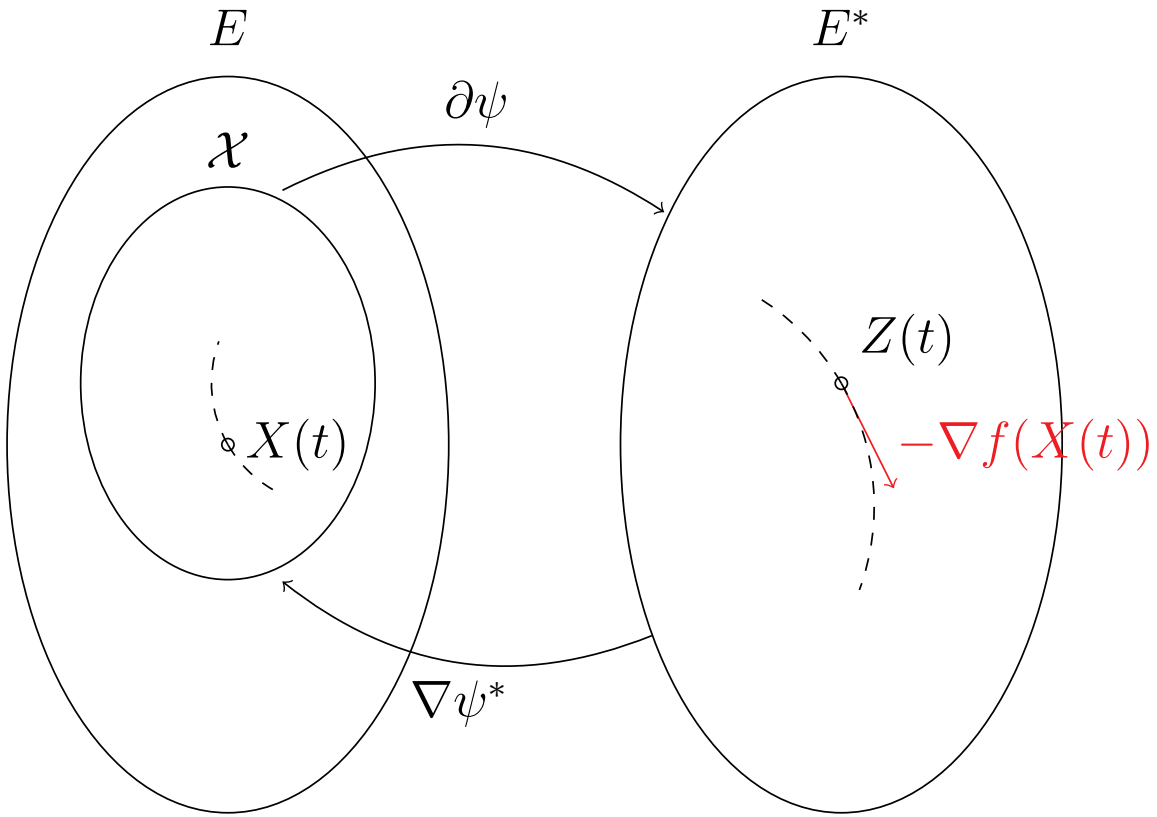

Figure: Illustration of Mirror Descent

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. [7]

1. Nesterov's method is discretization of

$$\begin{cases} \ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0 \\ X(0) = x_0,\ \dot{X}(0) = 0 \end{cases}$$

[7]Weijie Su, Stephen Boyd, and Emmanuel 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. [7]

1. Nesterov's method is discretization of

$$\begin{cases} \ddot{X} + \frac{r+1}{t}\dot{X} + \nabla f(X) = 0 \\ X(0) = x_0, \ \dot{X}(0) = 0 \end{cases}$$

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

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

---

[7]Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *NIPS*, 2014

Accelerated Mirror Descent (AMD) in Continuous Time

Lyapunov function

We start from a Lyapunov function.

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

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

Lyapunov function

We start from a Lyapunov function.

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

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

AMD ODE

If (X, Z) is a solution to

$$\begin{cases} \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t}{r}\nabla f(X), \\ X(0) = x_0, \ Z(0) = z_0, \text{ with } \nabla\psi^*(z_0) = x_0. \end{cases} \tag{1}$$

Then V is a Lyapunov function.

Convergence rate

Consequence:

Convergence rate

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

Proof:

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

Averaging Interpretation

$$\begin{cases} \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t}{r}\nabla f(X), \\ X(0) = x_0, \ Z(0) = z_0, \ \text{with } \nabla\psi^*(z_0) = x_0. \end{cases}$$

Averaging interpretation

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

Averaging Interpretation

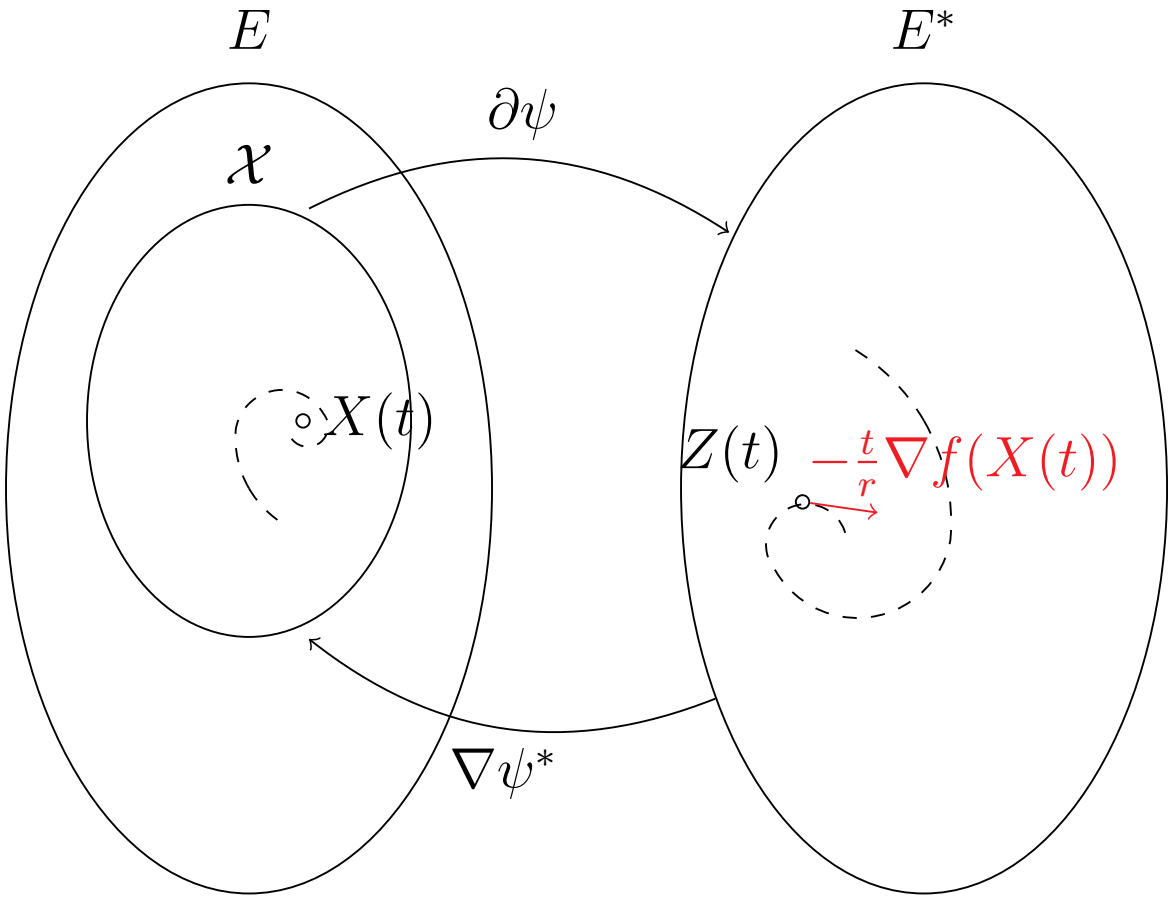

Figure: Averaging interpretation: Z evolves in E^*, X is a weighted average of the mirrored trajectory $\nabla\psi^*(Z)$.

Existence and uniqueness of the solution

Solution

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

Existence and uniqueness of the solution

Proof sketch (mostly from [7]):

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

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

Existence and uniqueness of the solution

Proof sketch (mostly from [7]):

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

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

1. Define family of "smoothed" ODEs:

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

Existence and uniqueness of the solution

Proof sketch (mostly from [7]):

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

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

1. Define family of "smoothed" ODEs:

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

2. Show that the solution (X_δ, Z_δ), $\delta = 2^{-i}$ is equi-Lipschitz continuous on $[0, t_0]$.

Existence and uniqueness of the solution

Proof sketch (mostly from [7]):

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

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

1. Define family of "smoothed" ODEs:

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

2. Show that the solution (X_δ, Z_δ), $\delta = 2^{-i}$ is equi-Lipschitz continuous on $[0, t_0]$.
3. By Arzelà-Ascoli theorem, there exists a converging subsequence. Its limit is a solution to (1).

Summary

So far:

- Family of accelerated ODEs.
- Existence and uniqueness of the solution.
- Converges at a $\mathcal{O}(1/t^2)$ rate.

AMD in Discrete Time

Discretization

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

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

First attempt:

$$\begin{cases} \frac{x^{(k+\mathbf{1})} - x^{(k)}}{\sqrt{s}} = \frac{r}{k\sqrt{s}}\left(\nabla\psi^*(z^{(k)}) - x^{(k+1)}\right), \\ \frac{z^{(k+\mathbf{1})} - z^{(k)}}{\sqrt{s}} + \frac{k\sqrt{s}}{r}\nabla f(x^{(k+1)}) = 0. \end{cases} \tag{2}$$

Discretization

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

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

First attempt:

$$\begin{cases} \frac{x^{(k+1)} - x^{(k)}}{\sqrt{s}} = \frac{r}{k\sqrt{s}}\left(\nabla\psi^*(z^{(k)}) - x^{(k+1)}\right), \\ \frac{z^{(k+1)} - z^{(k)}}{\sqrt{s}} + \frac{k\sqrt{s}}{r}\nabla f(x^{(k+1)}) = 0. \end{cases} \tag{2}$$

Equivalently,

$$\begin{cases} x^{(k+1)} = \lambda_k \nabla\psi^*(z^{(k)}) + (1-\lambda_k)x^{(k)}, \ \lambda_k = \frac{r}{r+k}, \\ z^{(k+1)} = z^{(k)} - \frac{ks}{r}\nabla f(x^{(k+1)}). \end{cases} \tag{3}$$

Candidate Lyapunov function

Candidate Lyapunov function:

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

Can show that

Comparison to the continuous case

$$\frac{d}{dt}V(X(t), Z(t), t) \leq -t\frac{r-2}{r}(f(X) - f^\star)$$

$$E^{(k+1)} - E^{(k)} \leq -\frac{s[(r-2)k-1]}{r}(f(x^{(k+1)}) - f^\star) + rD_{\psi^*}(z^{(k+1)}, z^{(k)}).$$

Extra term $rD_{\psi^*}(z^{(k+1)}, z^{(k)})$.

A small modification

Accelerated mirror descent with distance generating function ψ^*, regularizer R, step size s, and parameter $r \geq 3$

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

- R regularizer function, assumed ℓ_R strongly convex and L_R smooth.

A small modification

Accelerated mirror descent with distance generating function ψ^*, regularizer R, step size s, and parameter $r \geq 3$

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

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

The mirror operator $\nabla\psi^*$

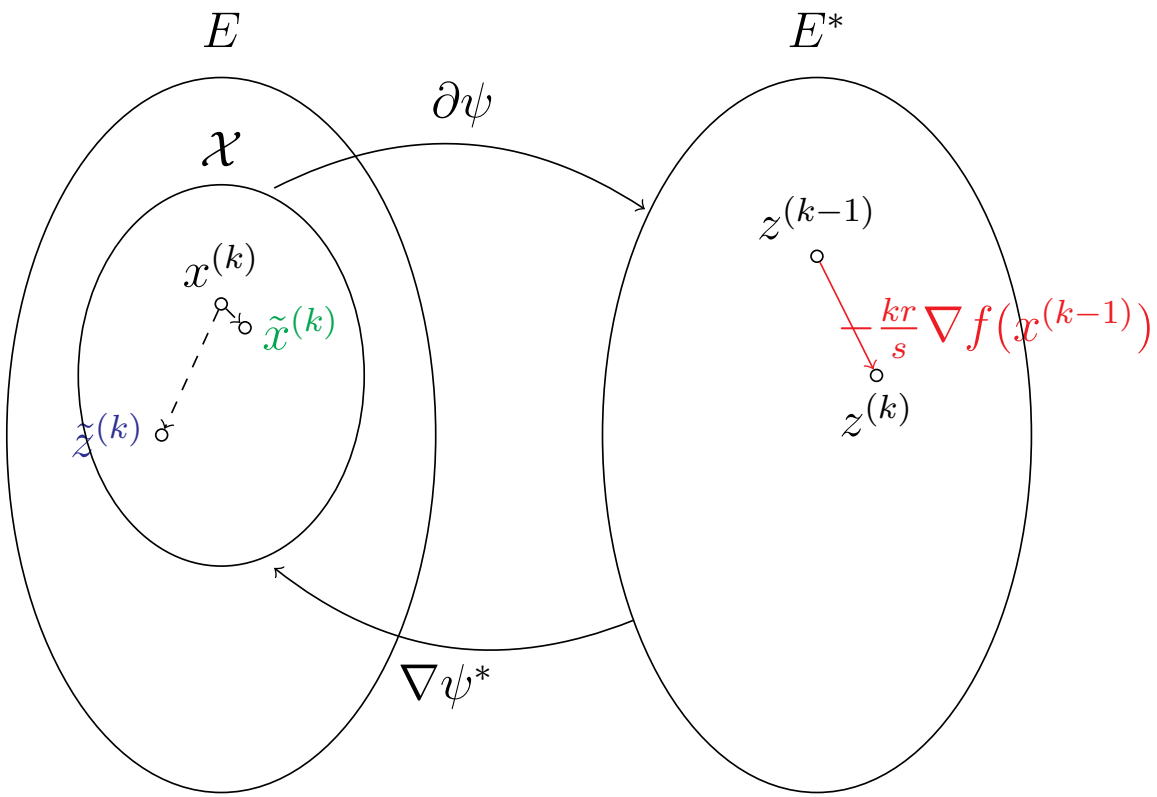

Figure: Illustration of the discrete AMD algorithm.

Convergence rate

$$\tilde{E}^{(k)} = V(\tilde{x}^{(k)}, z^{(k)}, k\sqrt{s}) = \frac{k^2 s}{r}(f(\tilde{x}^{(k)}) - f^\star) + rD_{\psi^*}(z^{(k)}, z^\star)$$

is a Lyapunov function under the following conditions:

- $\gamma \geq L_E L_{\psi^*}$
- $s \leq \frac{\ell_R}{2L_f\gamma}$

Convergence rate

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

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

Example: accelerated entropic descent on the simplex

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

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

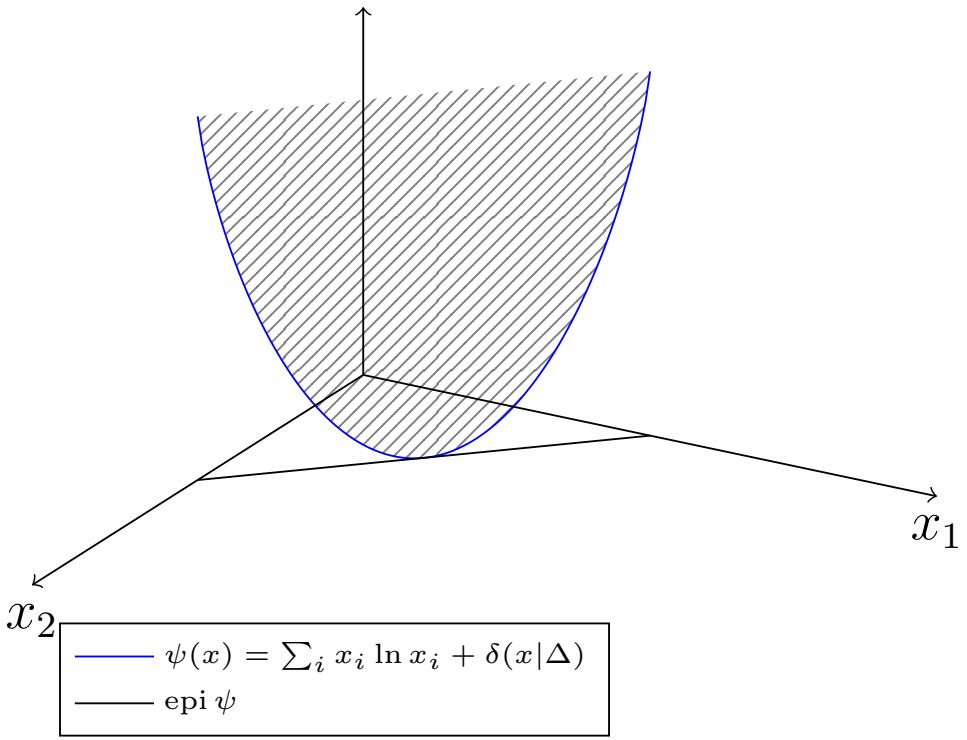

Figure: Illustration in $\mathbb{R}^2$.

(For the $\tilde{x}$ step, we take a "smoothed" entropy $\phi(x) = \psi(x + \epsilon)$)

Example: accelerated entropic descent on the simplex

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

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

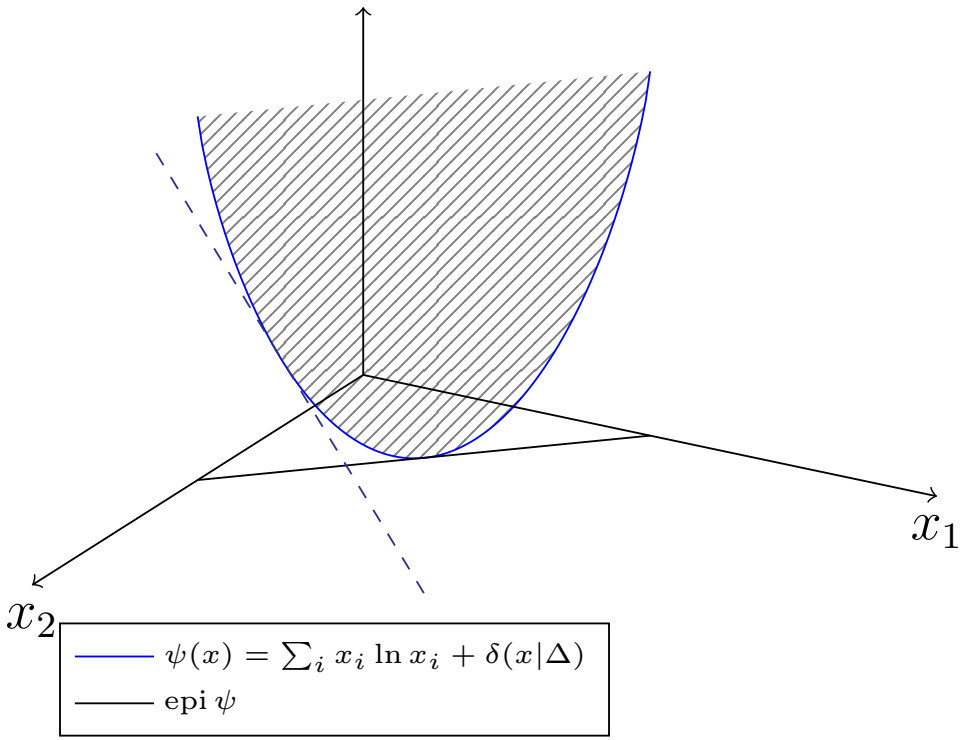

Figure: Illustration in $\mathbb{R}^2$.

(For the $\tilde{x}$ step, we take a "smoothed" entropy $\phi(x) = \psi(x + \epsilon)$)

Example: accelerated entropic descent on the simplex

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

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

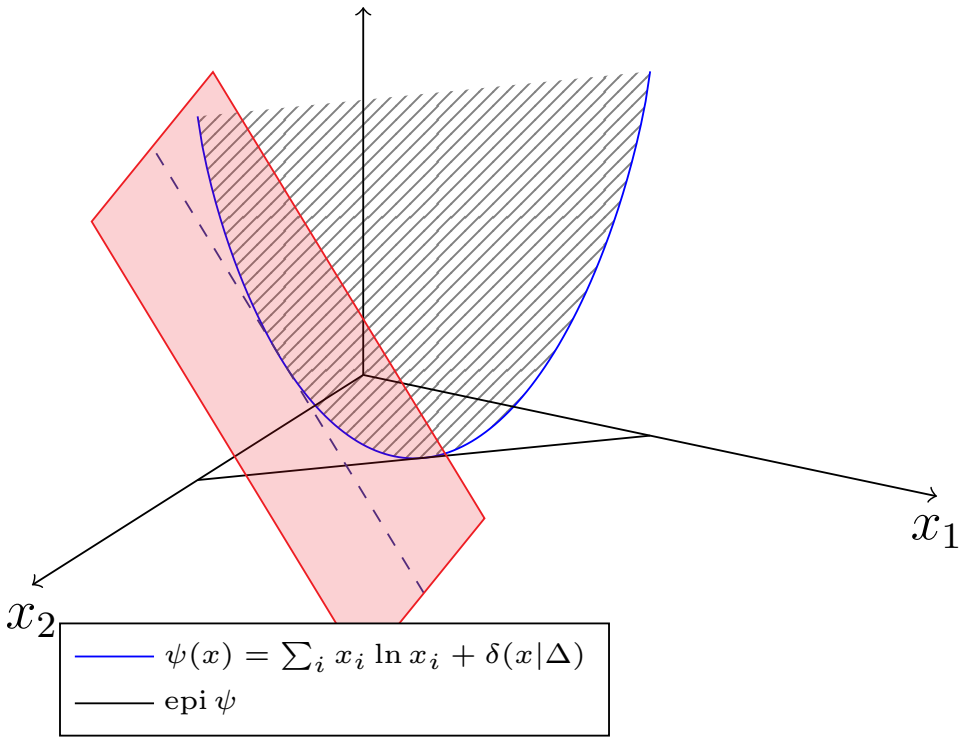

Figure: Illustration in $\mathbb{R}^2$.

(For the $\tilde{x}$ step, we take a "smoothed" entropy $\phi(x) = \psi(x + \epsilon)$)

Example: accelerated entropic descent on the simplex

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

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

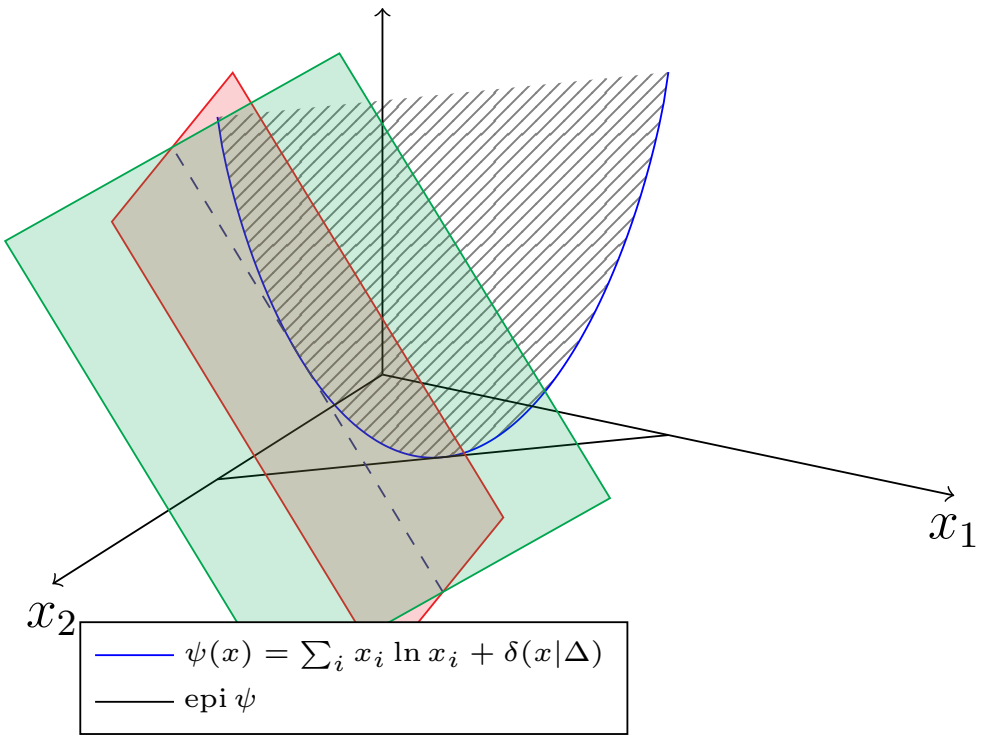

Figure: Illustration in $\mathbb{R}^2$.

(For the $\tilde{x}$ step, we take a "smoothed" entropy $\phi(x) = \psi(x + \epsilon)$)

Example: accelerated entropic descent on the simplex

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

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

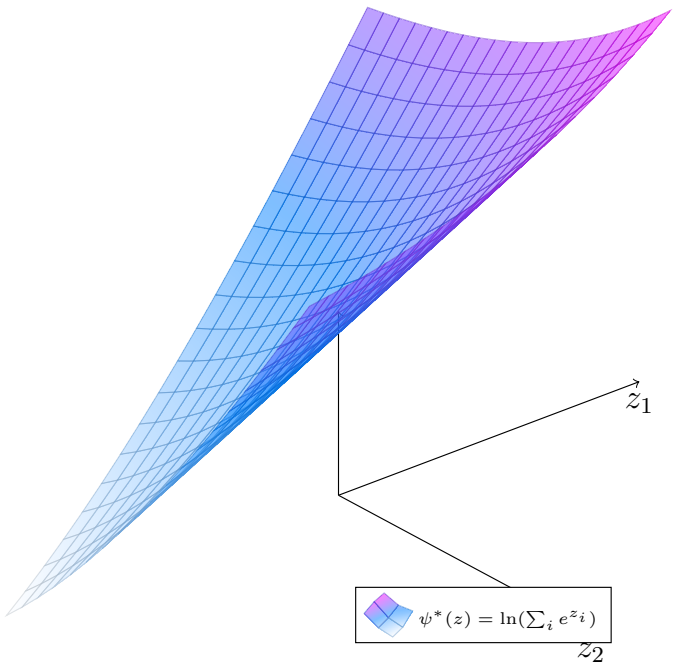

Figure: Illustration in $\mathbb{R}^2$.

(For the $\tilde{x}$ step, we take a "smoothed" entropy $\phi(x) = \psi(x + \epsilon)$)

Example: accelerated entropic descent on the simplex

Accelerated entropic descent.

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

Numerical Experiments and Restart Heuristics

Numerical experiments

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

Restarting

Restart the algorithm when a certain condition holds.

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

Restarting in continuous time

Illustration of restarting

Figure: Illustration of restarting

Effect of the parameter r

Figure: Effect of the parameter r, quadratic example

Effect of the parameter r

Empirically, r seems to reduce the period of oscillation. Leads to

- Slower progress for small k
- Faster progress for large k

Example with $x^\star \in \partial\Delta$

Figure: Example with $x^\star \in \partial\Delta$

Example with a non-smooth function

Figure: Non-smooth function, minimization of $f(x) = \|x\|$ on the simplex.

Summary

With appropriate choice of Lyapunov function, designed an ODE for Accelerated Mirror Descent

$$\begin{cases} \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t}{r}\nabla f(X), \\ X(0) = x_0,\ Z(0) = z_0,\ \text{with } \nabla\psi^*(z_0) = x_0. \end{cases}$$

- Existence and uniqueness of solution in continuous time
- Converges in $\mathcal{O}(1/t^2)$
- Averaging interpretation
- Discretization converges in $\mathcal{O}(1/k^2)$

Open problems

- Convergence of orbits $X(t)$?
  We have
  $$f(X(t)) - f^\star = \mathcal{O}(1/t^2)$$
  If $\mathcal{X}$ is compact, by continuity of f, $d(X(t), \mathcal{X}^*) \to 0$. But do we have convergence of $X(t)$ to some $x^\star \in \mathcal{X}^*$?

Open problems

- Convergence of orbits $X(t)$?
  We have
  $$f(X(t)) - f^\star = \mathcal{O}(1/t^2)$$
  If $\mathcal{X}$ is compact, by continuity of f, $d(X(t), \mathcal{X}^\star) \to 0$. But do we have convergence of $X(t)$ to some $x^\star \in \mathcal{X}^\star$?
- More general averaging: we showed the averaging interpretation
  $$X(t) = \frac{\int_0^t w(\tau) \nabla \psi^*(Z(\tau)) d\tau}{\int_0^t w(\tau) d\tau}$$
  with $w(\tau) = \tau^{r-1}$.
  Can we prove the same rate with different weights?

Open problems

- Composite optimization

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

where ∇f is Lipschitz and g is a general convex function, e.g. ℓ_1 norm.

Open problems

- Composite optimization

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

  where ∇f is Lipschitz and g is a general convex function, e.g. ℓ_1 norm.
- Rigorous analysis of effect of r.

Open problems

- Composite optimization

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

  where ∇f is Lipschitz and g is a general convex function, e.g. ℓ_1 norm.
- Rigorous analysis of effect of r.
- Prove restarted ODE converges faster (for strongly convex functions).

Open problems

- Composite optimization

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

  where ∇f is Lipschitz and g is a general convex function, e.g. ℓ_1 norm.
- Rigorous analysis of effect of r.
- Prove restarted ODE converges faster (for strongly convex functions).
- Design mirror operators for simple convex sets (e.g. polytopes).
  Desired properties:
  - Effective domain of ψ is $\mathcal{X}$.
  - ψ is cofinite and strongly convex.
  - $\nabla \psi^*$ is easy to compute.

Thank you!

References I

[1] Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956.

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

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

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

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

[6] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[7] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *NIPS*, 2014.

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

Back to mirror descent

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

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

[6]R.T. Rockafellar. *Convex Analysis*.
Princeton University Press, 1970

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

Back to mirror descent

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

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  ($\operatorname{dom} \psi^* = E^* \Leftrightarrow \operatorname{epi} \psi$ does not contain non-vertical half lines)

[6]R.T. Rockafellar. *Convex Analysis*.
Princeton University Press, 1970

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

Back to mirror descent

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

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  ($\operatorname{dom} \psi^* = E^* \Leftrightarrow \operatorname{epi} \psi$ does not contain non-vertical half lines)
- $\partial\psi^*(z) = \arg\max_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  (so $\partial\psi^*$ naturally maps into $\mathcal{X}$).

---

[6]R.T. Rockafellar. *Convex Analysis*.
Princeton University Press, 1970

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

Back to mirror descent

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

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  ($\operatorname{dom} \psi^* = E^* \Leftrightarrow \operatorname{epi} \psi$ does not contain non-vertical half lines)
- $\partial\psi^*(z) = \arg\max_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  (so $\partial\psi^*$ naturally maps into $\mathcal{X}$).
- ψ^* is (essentially) differentiable iff ψ is (essentially) strongly convex.

(these facts can be found in [6])

---

[6]R.T. Rockafellar. *Convex Analysis*.
Princeton University Press, 1970

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

Back to mirror descent

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

- $\psi : \mathcal{X} \to \mathbb{R}$
- $\psi^* : E^* \to \mathbb{R}$, $\psi^*(z) = \sup_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  ($\operatorname{dom} \psi^* = E^* \Leftrightarrow \operatorname{epi} \psi$ does not contain non-vertical half lines)
- $\partial\psi^*(z) = \arg\max_{x \in \mathcal{X}} \langle z, x \rangle - \psi(x)$
  (so $\partial\psi^*$ naturally maps into $\mathcal{X}$).
- ψ^* is (essentially) differentiable iff ψ is (essentially) strongly convex.

(these facts can be found in [6])

Mirror operator

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

---

[6]R.T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970

The mirror operator $\nabla\psi^*$

Back to mirror descent

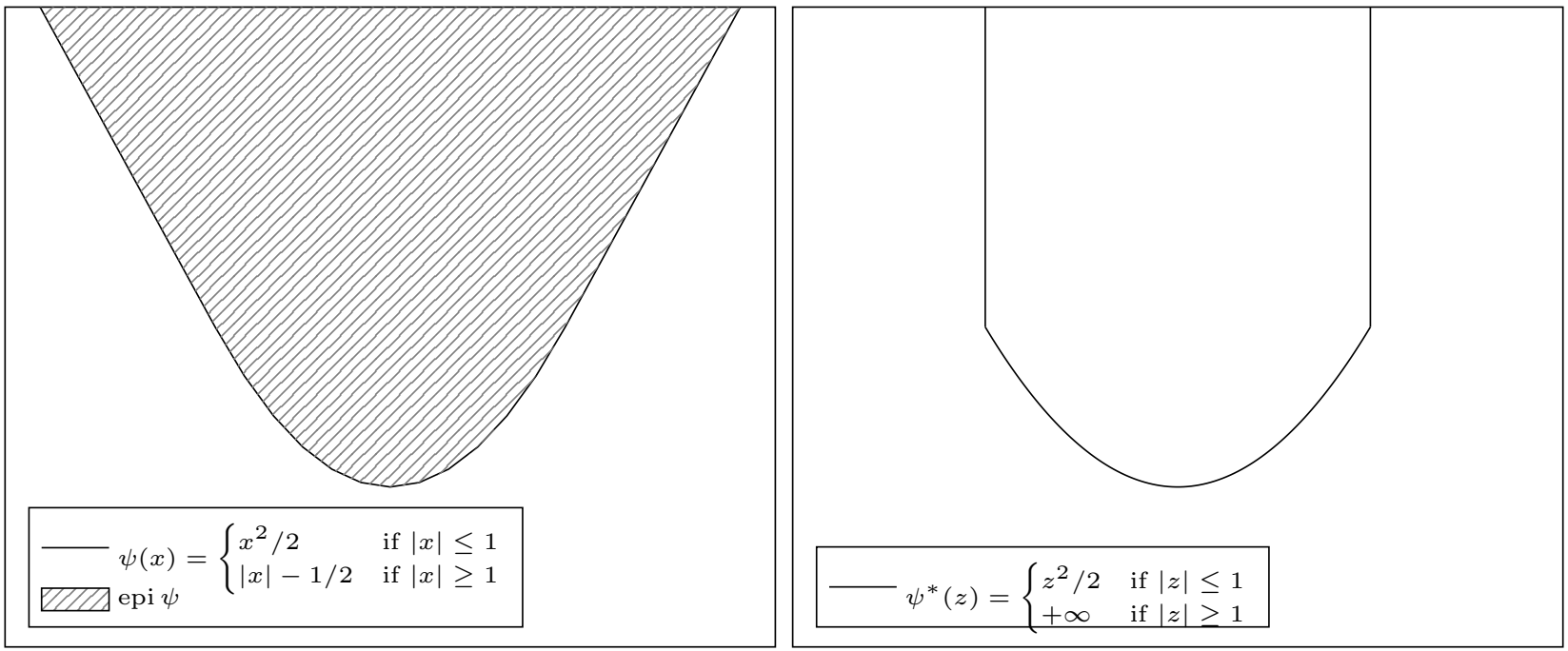

Figure: Example of dual distance generating functions ψ and ψ^*.

Restarting the ODE

Back Restarting the ODE at t_0 means

$$\begin{cases} \dot{X} = \frac{r}{t-t_0}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t-t_0}{r}\nabla f(X), \\ X(t_0) = X(t_0),\ Z(t_0) = z_0,\ \text{with } \nabla\psi^*(z_0) = X(t_0). \end{cases}$$

---

[5]Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes.
Foundations of Computational Mathematics, pages 1–18, 2013

Restarting the ODE

Back Restarting the ODE at t_0 means

$$\begin{cases} \dot{X} = \frac{r}{t-t_0}(\nabla\psi^*(Z) - X), \\ \dot{Z} = -\frac{t-t_0}{r}\nabla f(X), \\ X(t_0) = X(t_0),\ Z(t_0) = z_0,\ \text{with } \nabla\psi^*(z_0) = X(t_0). \end{cases}$$

Why restart?

- Some heuristics are known to (empirically) improve convergence [5].
  Intuitively: restart when X moves in the "wrong direction".
  E.g. $\left\langle \nabla f(X), \dot{X} \right\rangle \geq 0$.
- In the strongly convex case.

---

[5]Brendan O'Donoghue and Emmanuel Candès. Adaptive restart for accelerated gradient schemes.
Foundations of Computational Mathematics, pages 1–18, 2013

Restarting the ODE

Back

Suppose

- f is strongly convex: $\frac{\ell_f}{2}\|x - x^\star\|^2 \leq f(x) - f^\star$
- ψ^* is strongly convex: $\frac{\ell_{\psi^*}}{2}\|z - z^\star\|_*^2 \leq D_{\psi^*}(z, z^\star) \leq \frac{1}{2\ell_{\psi^*}}\|x - x^\star\|^2$

We restart every T.

Restarting the ODE

Back

Suppose

- f is strongly convex: $\frac{\ell_f}{2}\|x - x^\star\|^2 \leq f(x) - f^\star$
- ψ^* is strongly convex: $\frac{\ell_{\psi^*}}{2}\|z - z^\star\|_*^2 \leq D_{\psi^*}(z, z^\star) \leq \frac{1}{2\ell_{\psi^*}}\|x - x^\star\|^2$

We restart every T.

$$\begin{aligned} f(X((k+1)T)) - f^\star &\leq \frac{r^2 D_{\psi^*}(Z(kT), z^\star)}{T^2} \\ &\leq \frac{2r^2}{\ell_{\psi^*} T^2}\|X(kT) - x^\star\|^2 \\ &\leq \frac{r^2}{\ell_{\psi^*}\ell_f T^2}(f(X(kT)) - f^\star) \end{aligned}$$

Every epoch T, distance to optimum decreases by a constant $\frac{r^2}{\ell_{\psi^*}\ell_f T^2}$.