Accelerated Mirror Descent in Continuous and Discrete Time

Walid Krichene[†] Alexandre Bayen[†] Peter Bartlett[†*]

[†]Electrical Engineering and Computer Sciences, UC Berkeley, [*] Statistics, UC Berkeley, [*]Mathematical Sciences, QUT

{krichene, bayen, bartlett}@berkeley.edu

Continuous-time optimization

- Unconstrained smooth minimization, $\min_{x \in E} f(x)$, with ∇f Lipschitz.
- Optimization algorithms can be obtained by discretizing an ODE.

Gradient descent ODE

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

$$\frac{d}{dt}\frac{\|X - x^\star\|^2}{2} = \left\langle \dot{X}, X - x^\star \right\rangle = -\langle \nabla f(X), X - x^\star \rangle \le -(f(X) - f^\star)$$

Nesterov's ODE: Su, Boyd, Candès, (2014)

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

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

Accelerated Mirror descent

- Constrained smooth minimization, $\min_{x \in \mathcal{X}} f(x)$, with ∇f Lipschitz, $\mathcal{X}$ closed convex.
- Idea: Start with energy function, design ODE to make it a Lyapunov function.
- Mirror map: $\nabla \psi^* : E^* \to \mathcal{X}$.

Mirror descent: Nemirovski, Yudin (1983)

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

- Energy function on the dual space: Bregman divergence
  $D_{\psi^*}(Z, z^\star) = \psi^*(Z) - \psi^*(z^\star) - \langle \nabla \psi^*(z^\star), Z - z^\star \rangle$
  $$\frac{d}{dt}D_{\psi^*}(Z, z^\star) = \left\langle \dot{Z}, \nabla\psi^*(Z) - \nabla\psi^*(z^\star) \right\rangle = -\langle \nabla f(X), X - x^\star \rangle \le -(f(X) - f^\star)$$

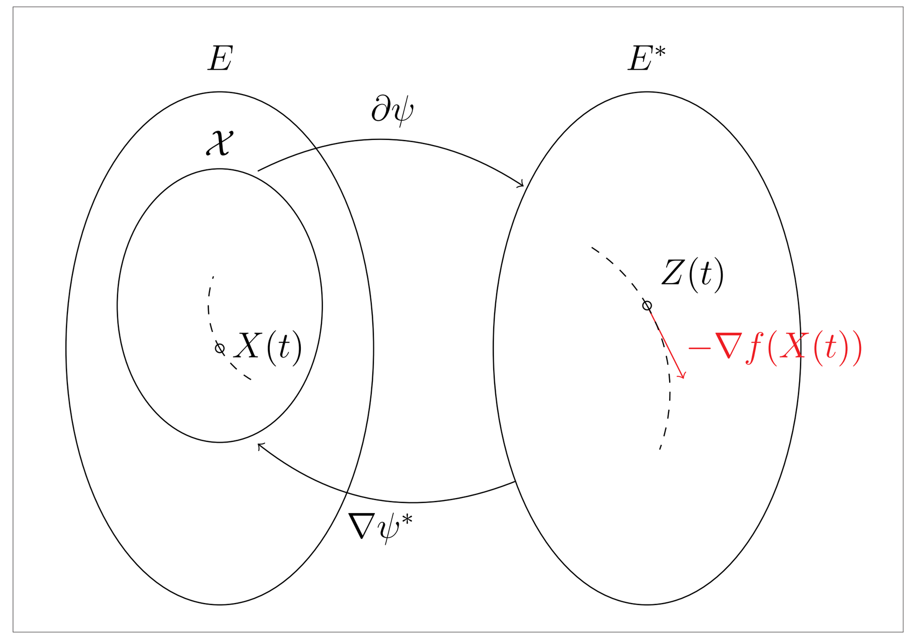


Figure: Mirror descent:
Dual: unconstrained gradient dynamics
Primal: mirror of dual.

Accelerated Mirror descent

$$\begin{cases} \dot{Z} = -\frac{t}{r}\nabla f(X) \\ \dot{X} = \frac{r}{t}(\nabla\psi^*(Z) - X) \\ X(0) = \nabla\psi^*(Z(0)) = x_0 \end{cases} \quad (1)$$

- Energy function $V(X, Z, t) = \frac{t^2}{r^2}(f(X) - f^\star) + D_{\psi^*}(Z, z^\star)$
  $$\begin{aligned} \frac{d}{dt}V(X(t), Z(t), t) &= \frac{2t}{r^2}(f(X) - f^\star) + \frac{t^2}{r^2}\left\langle \nabla f(X), \dot{X} \right\rangle + \left\langle \dot{Z}, \nabla\psi^*(Z) - \nabla\psi^*(z^\star) \right\rangle \\ &= \frac{2t}{r^2}(f(X) - f^\star) - \frac{t}{r}\langle \nabla f(X), X - x^\star \rangle \\ &\le -t\frac{r-2}{r^2}(f(X) - f^\star) \end{aligned}$$
- Averaging interpretation:
  $X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}$, $w(\tau) = \tau^{r-1}$

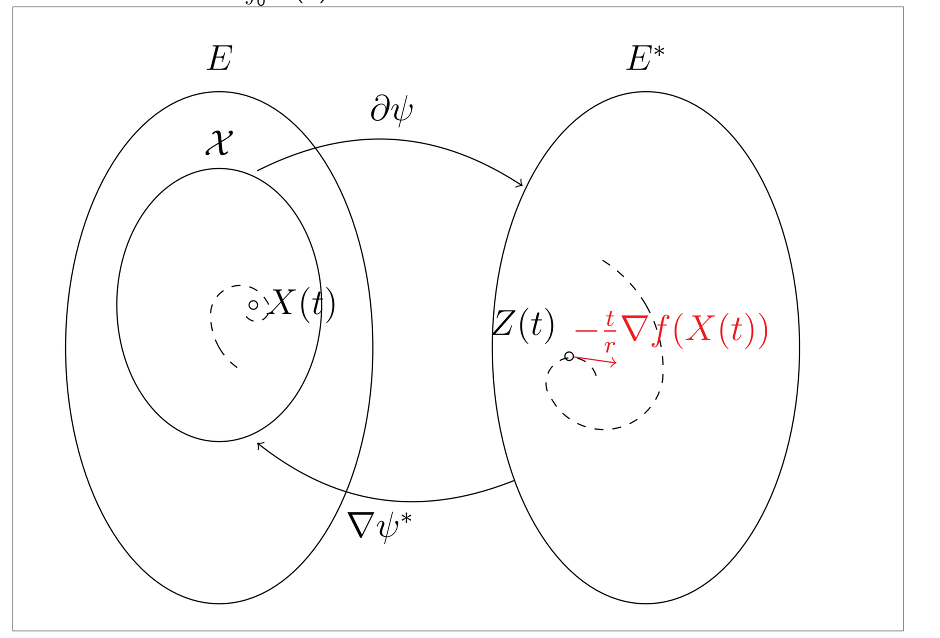


Figure: Accelerated Mirror descent:
Dual: unconstrained gradient dynamics with increasing weights.
Primal: average of mirrored dual.

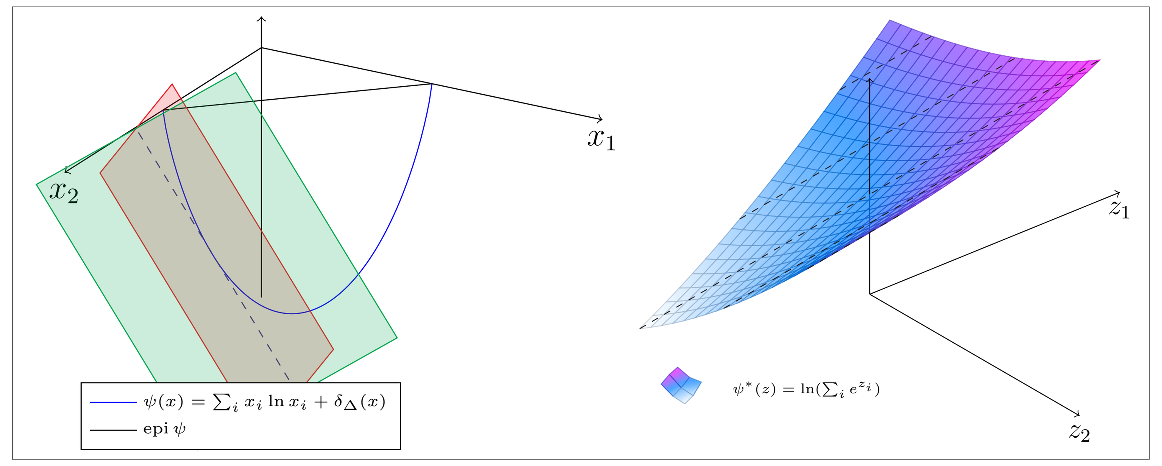
Mirror map

ψ, ψ^* conjugate pair of closed convex proper functions.

Proposition

If ψ is strictly convex, $\mathrm{dom}\,\psi = \mathcal{X}$ and $\mathrm{epi}\,\psi$ has trivial recession cone, then ψ^* is finite and differentiable on E^*, and $\nabla\psi^* : E^* \to \mathcal{X}$.

- Note: ψ not necessarily differentiable.
- Examples on $\mathcal{X} = \Delta$ the n-simplex:
- Entropy:
  $\psi(x) = \sum_i x_i \ln x_i + \delta_\Delta(x)$
  $\psi^*(z) = \ln \sum_i e^{z_i}$ and $\nabla\psi^*(z)_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$
- Smoothed entropy:
  $\psi(x) = \sum_i (x_i + \epsilon)\ln(x_i + \epsilon) + \delta_\Delta(x)$
  $\nabla\psi^*(z)$ can be computed in $\mathcal{O}(n \log n)$.



Solutions and convergence rate

Theorem

If ∇f and $\nabla\psi^*$ are Lipschitz, then ODE (1) has a unique solution, and the solution remains in $\mathcal{X}$.

- Consider smoothed ODE $\dot{X} = \frac{r}{\max(t,\delta)}(\nabla\psi^*(Z) - X)$. Has unique solution (X_δ, Z_δ) by Cauchy-Lipschitz.
- Extract a converging subsequence of $(X_{\delta_i}, Z_{\delta_i})_i$: Limit is solution to (1)

Theorem

If $r \ge 2$, then V is a Lyapunov function, and $f(X(t)) - f^\star \le \frac{r^2 D_{\psi^*}(Z(0), z^\star)}{t^2}$.

$$\begin{aligned} f(X(t)) - f^\star &\le \frac{r^2}{t^2}V(X(t), Z(t), t) \\ &\le \frac{r^2}{t^2}V(X(0), Z(0), 0) = \frac{r^2}{t^2}D_{\psi^*}(Z(0), z^\star) \end{aligned}$$

Discretization

Algorithm 1 Accelerated mirror descent in discrete time

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

- $\lambda_k = \frac{k}{k+r}$.
- R: ℓ_R strongly convex and 1-smooth regularizer.
- s: step size.
- Consistent with ODE as $s \to 0$.

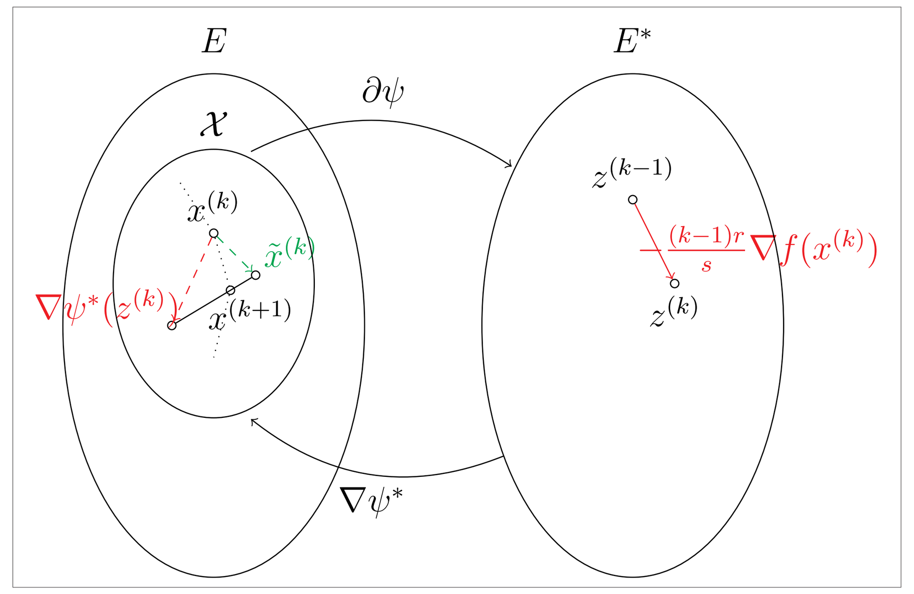


Figure: AMD in discrete time.

Theorem

If $s \le \frac{\ell_R}{2L_f L_{\psi^*}}$, then $f(\tilde{x}^{(k)}) - f^\star \le C/k^2$, where $C = \frac{r^2 D_{\psi^*}(z_0, z^\star)}{s} + f(x_0) - f^\star$.

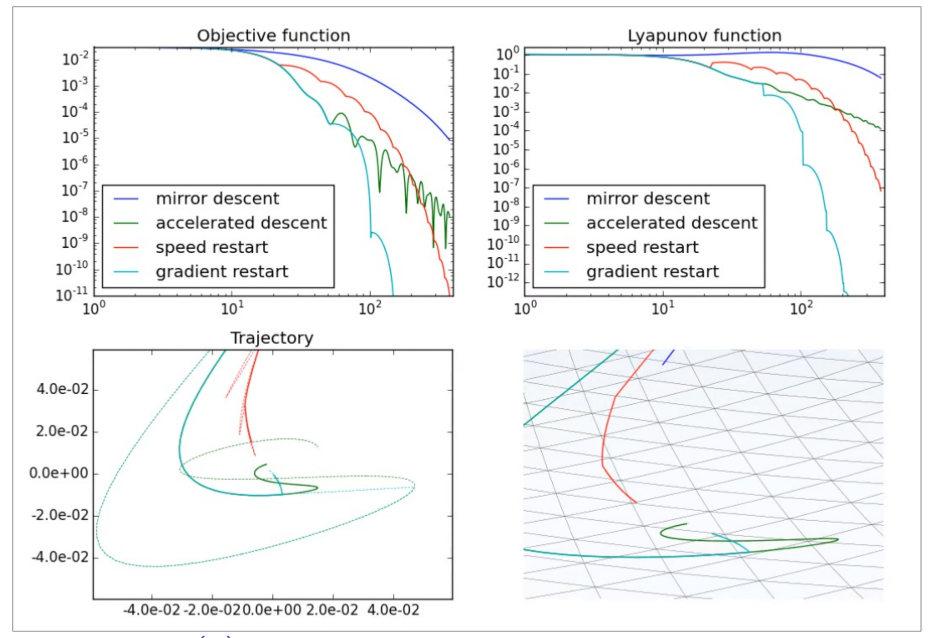
Proof: $V(\tilde{x}^{(k)}, z^{(k)}, k)$ is a Lyapunov function.

Restarting

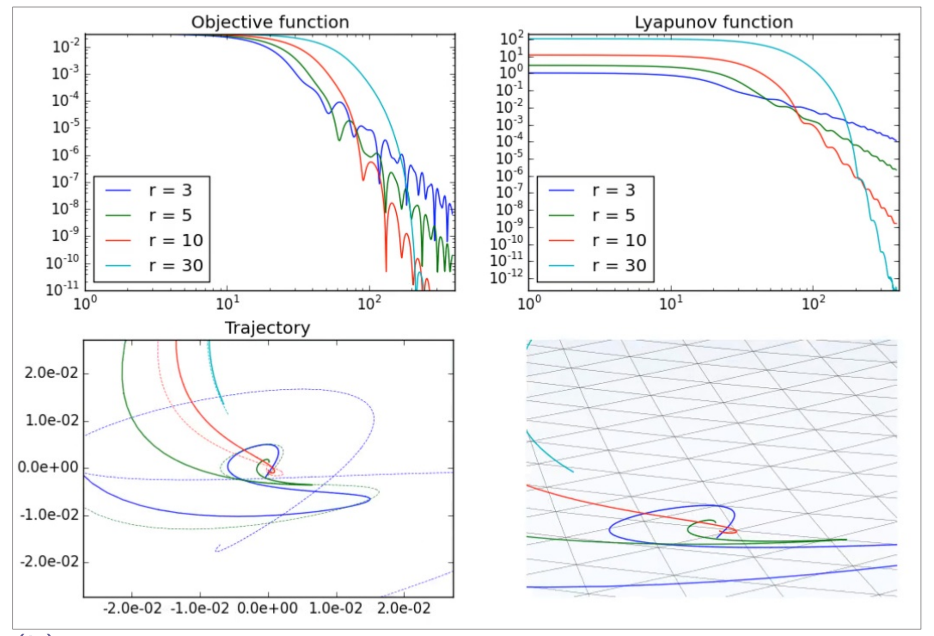
Restarting heuristics

"Reset time to 0" when
- $\left\langle \nabla f(x^{(k)}), x^{(k+1)} - x^{(k)} \right\rangle \ge 0$ (trajectory points in a bad direction)
- $\|x^{(k+1)} - x^{(k)}\| \le \|x^{(k)} - x^{(k-1)}\|$ (trajectory decelerates).

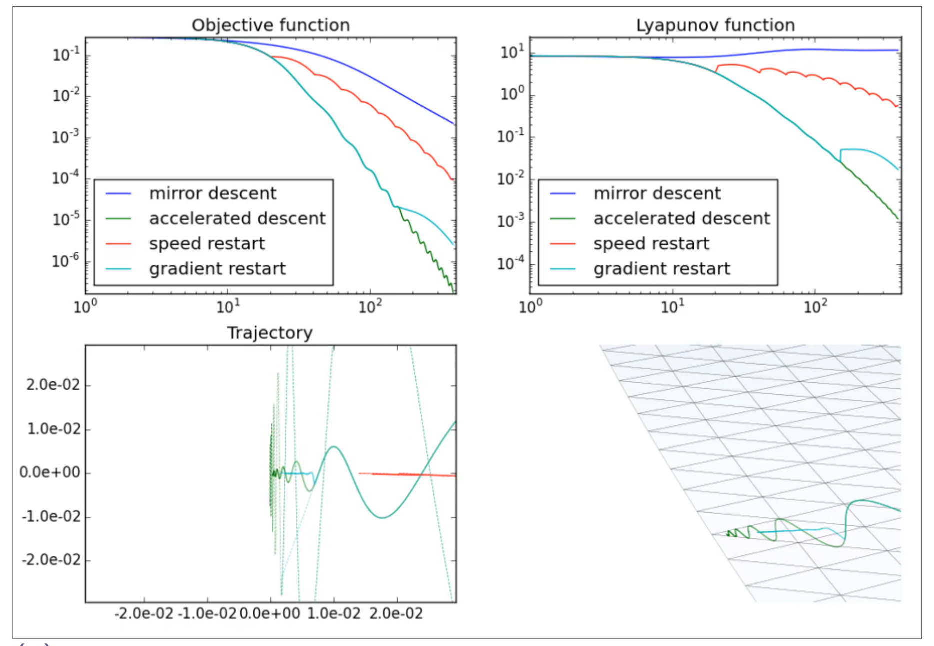
Numerical experiments



(a) Illustration of acceleration and restarting



(b) Effect of parameter r: Larger r is slower for small t and faster for large t; affects period of oscillations.



(c) Optimum on the boundary: Restarting may not always improve convergence rate.

Extensions, open questions

- Generalized averaging:
  $$X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}$$
- Primal representation:
  $$\ddot{X} + \frac{r+1}{t}\dot{X} = -\nabla^2\psi^* \circ \nabla\psi(X + \frac{t}{r}\dot{X})\nabla f(X)$$
  Constrained non-linear oscillator, vanishing damping.
- Prove convergence of trajectory $X(t)$.
- Prove faster rate for the restarted ODE.
- Adaptive choice of r?
- Accelerated ODE for composite optimization.

This work was supported in part by the NSF (CCF-1115788, CNS-1238959, CNS-1238962, CNS-1239054, CNS-1239166), the ARC (FL110100281 and ACEMS), and the Simons Institute Fall 2014 Algorithmic Spectral Graph Theory Program.