Adaptive Averaging in Accelerated Descent Dynamics

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

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

{walid, bayen, peter}@berkeley.edu

Accelerated Mirror Descent

- Optimization algorithms can be obtained by discretizing an ODE.
- 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

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

- Energy function: Bregman divergence

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

- Time derivative

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

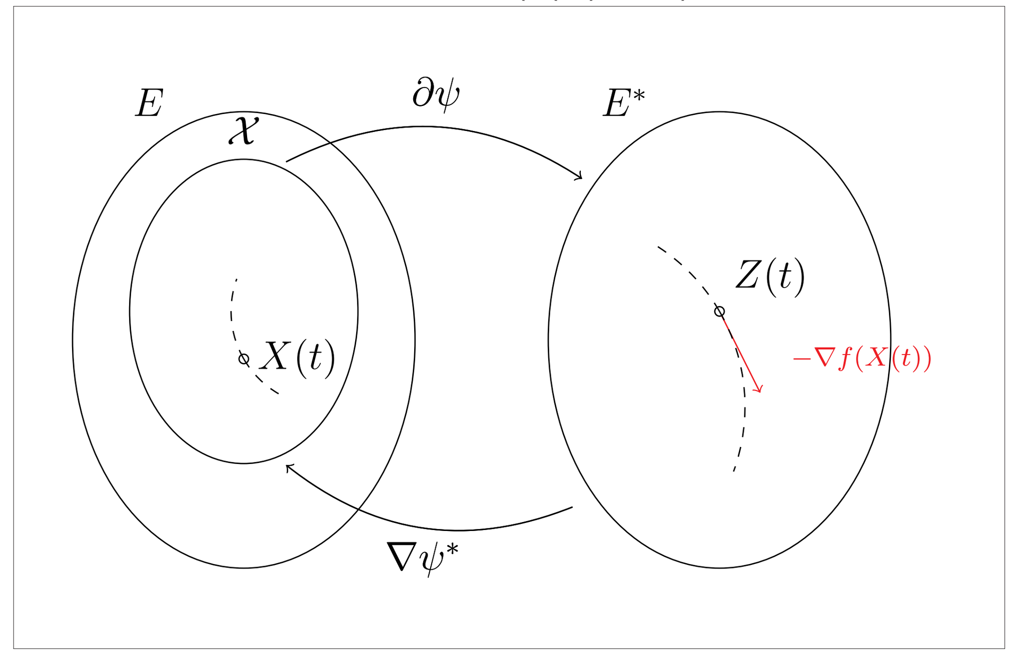


Figure: Mirror descent:
Dual: unconstrained gradient dynamics
Primal: mirror of dual.

Accelerated Mirror descent

$$\text{AMD}_{w,\eta}\begin{cases}\dot Z = -\eta(t)\nabla f(X),\\ X(t) = \frac{\int_0^t w(\tau)\nabla\psi^*(Z(\tau))d\tau}{\int_0^t w(\tau)d\tau}\\ \nabla\psi^*(Z(0)) = x_0\end{cases}$$

- Energy function

$$L_r(X, Z, t) = \underbrace{\boxed{r(t)(f(X) - f^*)}}_{\text{convergence rate}} + \underbrace{\boxed{D_{\psi^*}(Z, z^*)}}_{\text{constraints}}$$

- Time derivative

$$\frac{d}{dt}L_r(X(t), Z(t), t)$$
$$\le (f(X) - f^*)(r' - \eta) + \left\langle\nabla f(X), \dot X\right\rangle(r - \eta w/W)$$

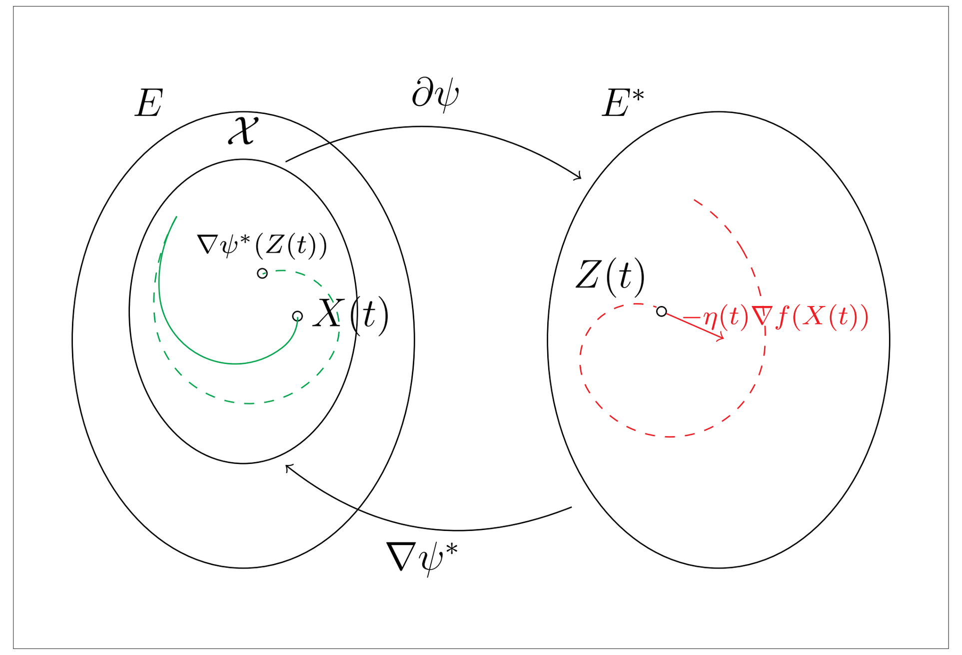


Figure: Accelerated Mirror Descent with Generalized Averaging:
Dual: unconstrained gradient dynamics with weight $\eta(t)$.
Primal: average of mirrored dual with weights $w(t)$.

Convergence rate

Theorem

Given a rate $r(t)$, if $w(t), \eta(t)$ satisfy $\left\langle\nabla f(X(t)), \dot X(t)\right\rangle\left(r(t) - \frac{\eta(t)w(t)}{W(t)}\right) \le 0$ and $\eta(t) \ge r'(t)$, then L_r is a Lyapunov function for $\text{AMD}_{w,\eta}$ and

$$f(X(t)) - f^* \le \frac{L_r(x_0, z_0, t_0)}{r(t)}$$

Furthermore, energy function $E_r(t) = f(X(t)) + \frac{1}{r(t)}D_{\psi^*}(Z(t), \dot X(t))$ is decreasing.

Re-parameterization: $a(t) = \frac{w(t)}{W(t)}$, $w(t) = w(t_0)\frac{a(t)}{a(t_0)}e^{\int_{t_0}^t a(\tau)d\tau}$.

A simple sufficient condition

Corollary

Given a rate $r(t)$, if $\eta(t) \ge r'(t)$ and $a(t) = \frac{\eta(t)}{r(t)}$, then L_r is a Lyapunov function for $\text{AMD}_{w,\eta}$ and

$$f(X(t)) - f^* \le \frac{L_r(x_0, z_0, t_0)}{r(t)}$$

Example: If $r(t) = t^p$, then we can take $\eta(t) = \beta t^{p-1}$, $\beta \ge p$ and $a(t) = \frac{\beta}{t}$ $(w(t) = (t/t_0)^{\beta-1})$.

Adaptive Averaging

Heuristic, guaranteed to preserve Lyapunov function L_r and energy E_r

$$\eta(t) \ge r'(t) \qquad \text{and} \qquad \begin{cases} a(t) = \frac{\eta(t)}{r(t)} & \text{if } \left\langle\nabla f(X(t)), \dot X(t)\right\rangle > 0,\\ a(t) \ge \frac{\eta(t)}{r(t)} & \text{Otherwise.}\end{cases}$$

Convergence guarantee in discrete time

Theorem

AMD with Adaptive Averaging, with step sizes $\gamma \ge \beta\beta^{\max}L_R L_{\psi^*}$ and $s \le \frac{\ell_R}{2L_f\gamma}$, guarantees

$$f(\tilde x^{(k)}) - f^* \le \frac{\tilde L^{(1)}}{sk^2}.$$

Algorithm 1 AMD with Adaptive Averaging

Initialize $\tilde x^{(0)} = x_0$, $\tilde z^{(0)} = x_0$, $a_1 = \frac{\beta}{\sqrt{s}}$
for $k \in \mathbb{N}$ **do**
$\tilde z^{(k+1)} = \arg\min_{\tilde z\in\mathcal{X}} \beta ks\left\langle\nabla f(x^{(k)}), \tilde z\right\rangle + D_\psi(\tilde z, \tilde z^{(k)})$.
$\tilde x^{(k+1)} = \arg\min_{\tilde x\in\mathcal{X}} \gamma s\left\langle\nabla f(x^{(k)}), \tilde x\right\rangle + R(\tilde x, x^{(k)})$
$x^{(k+1)} = \lambda_{k+1}\tilde z^{(k+1)} + (1 - \lambda_{k+1})\tilde x^{(k+1)}$, with $\lambda_k = \frac{\sqrt{s}a_k}{1+\sqrt{s}a_k}$.
$a_k = \min\left(a_{k-1}, \frac{\beta^{\max}}{k\sqrt{s}}\right)$
if $f(\tilde x^{(k+1)}) - f(\tilde x^{(k)}) > 0$ **then**
$a_k = \frac{\beta}{k\sqrt{s}}$

Example: Accelerated Replicator Dynamics on the simplex

An equivalent form in the primal space

Mirror descent (primal)

$$\text{MD}^p\begin{cases}\dot X = -\nabla^2\psi^*\circ\nabla\psi(X(t))\nabla f(X(t))\\ X(0) = x_0\end{cases}$$

Accelerated Mirror descent (primal)

$$\text{AMD}^p_{w,\eta}\begin{cases}\dot{\check Z} = -\eta(t)\nabla^2\psi^*\circ\nabla\psi(\check Z(t))\nabla f(X(t))\\ \check Z(0) = x_0\end{cases}$$

Simplex-constrained optimization:
$\psi(x) = \sum_i x_i \ln x_i + \delta_{\mathcal{X}}(x)$,
$\nabla^2\psi^*\circ\nabla\psi(x)_{i,j} = \delta^j_i x_i - x_i x_j$.

$$\text{Replicator}\begin{cases}\dot X_i = X_i(\langle X, \nabla f(X)\rangle - \nabla f(X)_i)\\ X(0) = x_0\end{cases}$$

$$\text{Acc. Rep.}\begin{cases}\dot{\check Z}_i = \eta(t)\check Z_i(\langle\check Z, \nabla f(X)\rangle - \nabla f(X)_i)\\ \check Z(0) = x_0\end{cases}$$

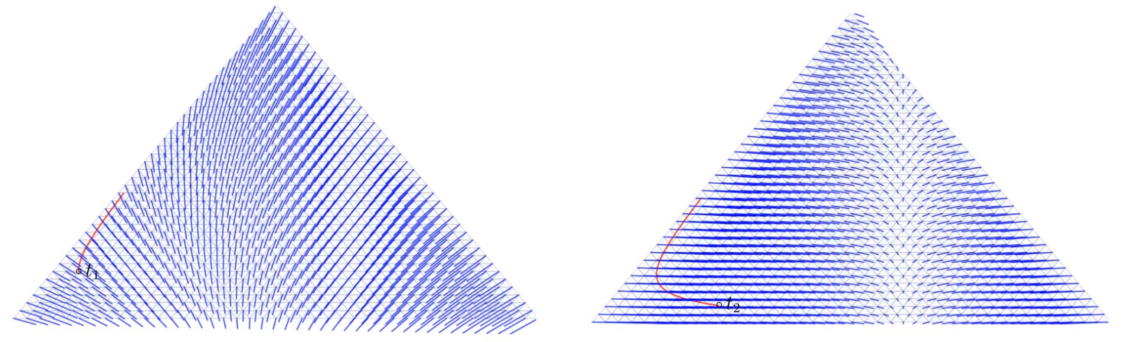
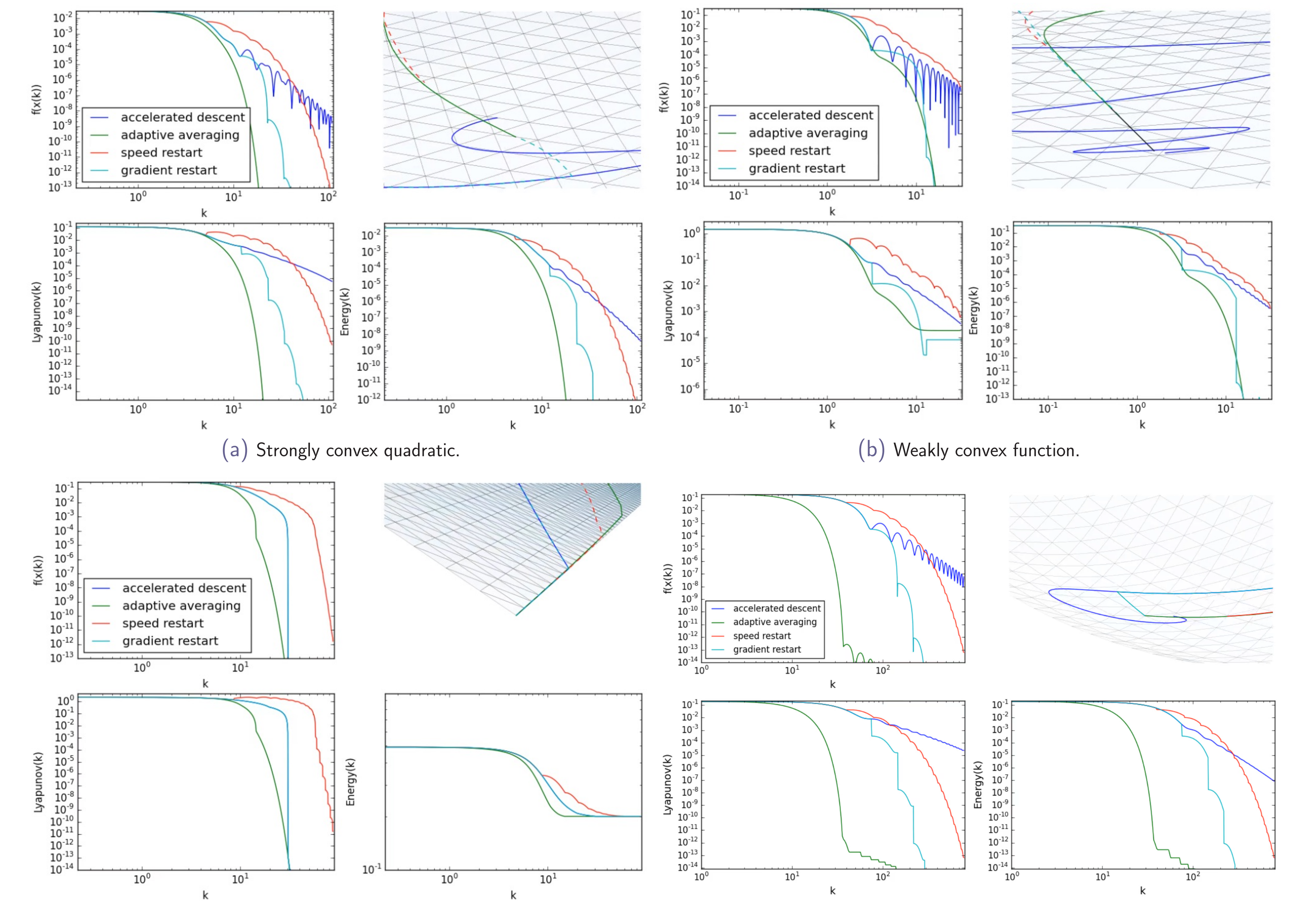


Figure: Vector field $X \mapsto \nabla^2\psi^*(Z(t))\nabla f(X)$ for different values of $Z(t)$.

Numerical experiments: Adaptive averaging and restarting



(a) Strongly convex quadratic.
(b) Weakly convex function.
(c) Linear function.
(d) KL divergence.

Summary of heuristics

Gradient Restart (O'Donoghue, Candès, 2015)	Speed Restart (Su, Boyd, Candès, 2014)	Adaptive Averaging
Damped non-linear oscillator	Damped non-linear oscillator	Generalized Averaging
Restart ($\dot X(t) = 0$) when $\langle\nabla f(X), \dot X\rangle > 0$	Restart ($\dot X(t) = 0$) when $\frac{d}{dt}\|\dot X\| < 0$	$a(t) = \begin{cases}\frac{\eta(t)}{r(t)} & \text{if } \langle\nabla f(X), \dot X\rangle > 0,\\ \text{const.} & \text{otherwise.}\end{cases}$
Restart when moving in bad direction	Restart when progress is slowing	Increase weights on good portions of trajectory

Extension

- Adaptive Averaging for higher-order accelerated methods: Cubic-regularized Newton method (Nesterov)