MATHEMATICS OF DEEP LEARNING

JOAN BRUNA , CIMS + CDS, NYU, SPRING'18

Lecture 7: Discrete vs Continuous time in Optimization: the convex case.

OBJECTIVES LECTURE 7

- Review of convex optimization
- Accelerated Gradient Methods
- Continuous vs Discrete time optimization

CONVEX OPTIMIZATION

- We consider $\min_{\theta} g(\theta)$ with $g : \mathbb{R}^d \to \mathbb{R}$ convex
- For generic (non-differentiable) g ,

$$\forall\ \theta_1, \theta_2, t \in [0,1]\ ,\ \ g(t\theta_1 + (1-t)\theta_2) \leq t g(\theta_1) + (1-t) g(\theta_2)\ .$$

CONVEX OPTIMIZATION

- We consider $\min_\theta g(\theta)$ with $g : \mathbb{R}^d \to \mathbb{R}$ convex
- For generic (non-differentiable) g ,

$$\forall\ \theta_1, \theta_2, t \in [0,1]\ ,\ \ g(t\theta_1 + (1-t)\theta_2) \leq tg(\theta_1) + (1-t)g(\theta_2)\ .$$

- Assuming g is differentiable,

$$\forall\ \theta_1, \theta_2, g(\theta_1) \geq g(\theta_2) + \langle \nabla g(\theta_2), \theta_1 - \theta_2 \rangle\ .$$

- In convex analysis, it is convenient to extend this property to non-differentiable functions.

SUBGRADIENTS

- Given $g : \mathbb{R}^d \to \mathbb{R}$ convex, $s \in \mathbb{R}^d$ is a *subgradient* of g at θ iff

$$\forall \theta' \in \mathbb{R}^d \ , \ g(\theta') \geq g(\theta) + \langle s, \theta' - \theta \rangle \ .$$

- Subdifferential:

$$\partial g(\theta) := \{ s \in \mathbb{R}^d ; s \text{ is a subgradient of } g \text{ at } \theta \}.$$

SUBGRADIENTS

- Given $g : \mathbb{R}^d \to \mathbb{R}$ convex, $s \in \mathbb{R}^d$ is a *subgradient* of g at θ iff

$$\forall \theta' \in \mathbb{R}^d \ , \ g(\theta') \geq g(\theta) + \langle s, \theta' - \theta \rangle \ .$$

- Subdifferential:

$$\partial g(\theta) := \{s \in \mathbb{R}^d ; s \text{ is a subgradient of } g \text{ at } \theta\}.$$

- It only contains the gradient when g is differentiable.
- It is never empty [Farkas lemma].

MAIN FUNCTIONAL ASSUMPTIONS

- *Lipschitz Continuity:* g is differentiable, convex and has gradients uniformly bounded on a compact set.

$$\|g'(\theta)\| \leq B \ \forall \ \theta; \|\theta\| \leq D$$

  - Equivalent to $|g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|$.

MAIN FUNCTIONAL ASSUMPTIONS

- *Lipschitz Continuity:* g is differentiable, convex and has gradients uniformly bounded on a compact set.

$$\|g'(\theta)\| \leq B \ \forall \ \theta; \|\theta\| \leq D$$

  - Equivalent to $|g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|$.

- *Smoothness:* g differentiable and its gradient is Lipschitz continuous:

$$\forall \ \theta_1, \theta_2 \ , \|g'(\theta_1) - g'(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$$

  - If g twice differentiable, $L\mathbf{I} \succeq \nabla^2 g(\theta)$

MAIN FUNCTIONAL ASSUMPTIONS

- *Lipschitz Continuity:* g is differentiable, convex and has gradients uniformly bounded on a compact set.

$$\|g'(\theta)\| \leq B \ \forall \ \theta; \|\theta\| \leq D$$

  - Equivalent to $|g(\theta) - g(\theta')| \leq B\|\theta - \theta'\|$.

- *Smoothness:* g differentiable and its gradient is Lipschitz continuous:

$$\forall \ \theta_1, \theta_2 \ , \|g'(\theta_1) - g'(\theta_2)\| \leq L\|\theta_1 - \theta_2\|$$

  - If g twice differentiable, $L\mathbf{I} \succeq \nabla^2 g(\theta)$

- *Strong convexity:*

$$\forall \ \theta_1, \theta_2 \ , g(\theta_1) \geq g(\theta_2) + \langle \nabla g(\theta_2), \theta_1 - \theta_2 \rangle + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2$$

  - If g twice differentiable, $\nabla^2 g(\theta) \succeq \mu\mathbf{I}$

CHARACTERIZATION OF CONVERGENCE

- Given an iterative scheme producing $(\theta_k)_{k=0,1,\dots}$, we are interested in several forms of convergence:
- Objective Function convergence:

$$l_k = g(\theta_k) - \inf_{\eta \in \mathbb{R}^d} g(\eta)$$

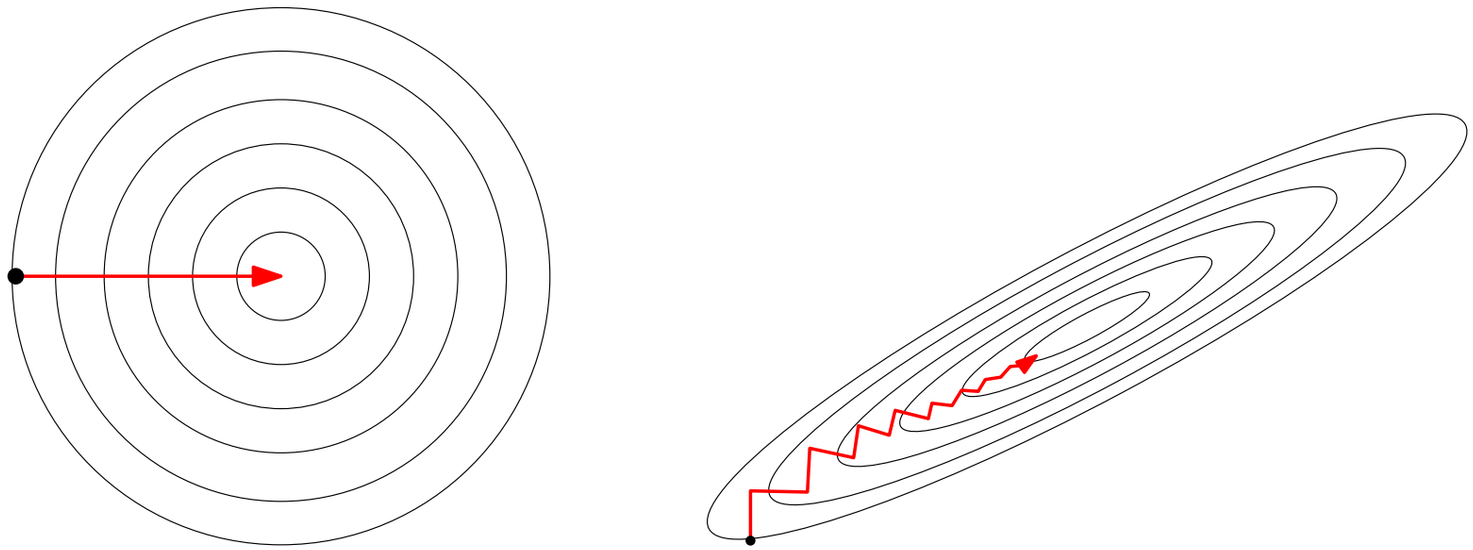
- Iterates convergence:

$$\inf_{\eta \in \arg\min g} \|\theta_k - \eta\|^2$$

GRADIENT DESCENT ON SMOOTH CONVEX FUNCTIONS

- Assume g is convex with L-Lipschitz continuous gradient.

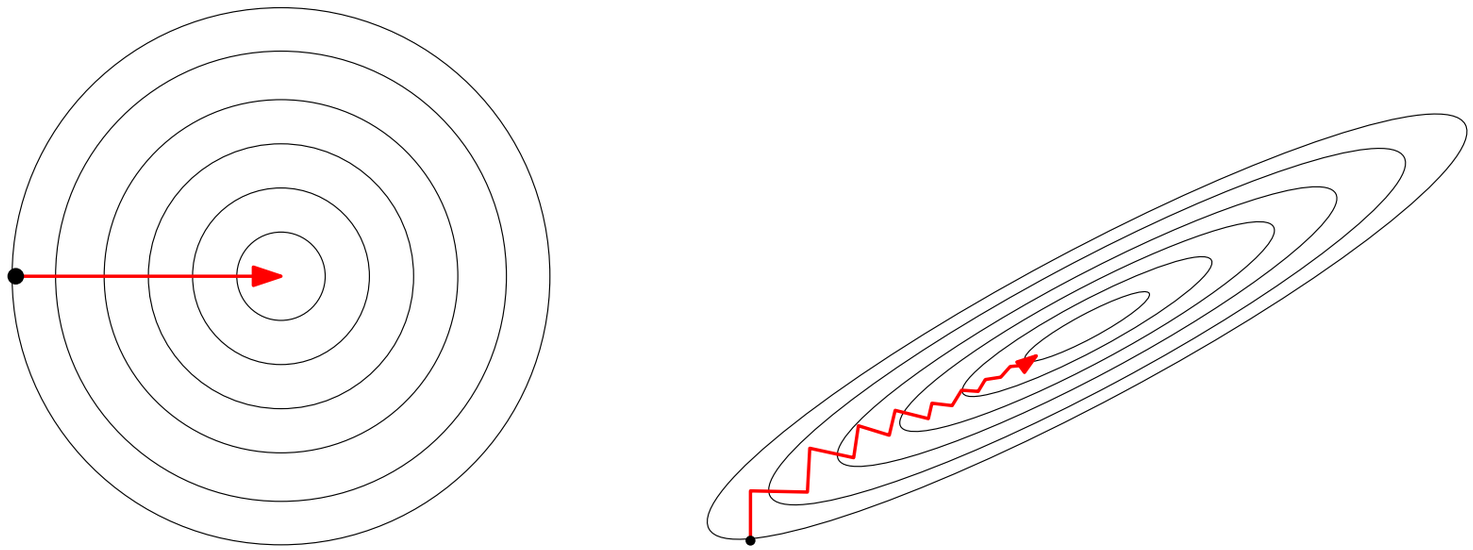
$$\theta_{k+1} = \theta_k - L^{-1} \nabla g(\theta_k).$$



GRADIENT DESCENT ON SMOOTH CONVEX FUNCTIONS

➤ Assume g is convex with L-Lipschitz continuous gradient.

$$\theta_{k+1} = \theta_k - L^{-1} \nabla g(\theta_k).$$



➤ Convergence Properties?

KEY PROPERTIES OF CONVEX FUNCTIONS

- Quadratic Approximation of L-smooth convex functions:

$$0 \leq g(\theta) - g(\beta) - \langle \nabla g(\beta), \theta - \beta \rangle \leq L \|\theta - \beta\|^2$$

- Co-coercivity for L-smooth convex functions:

$$\|\nabla g(\theta) - \nabla g(\beta)\|^2 \leq L \langle \nabla g(\theta) - \nabla g(\beta), \theta - \beta \rangle \ .$$

- For strongly-convex functions:

$$g(\theta) \leq g(\beta) + \langle \nabla g(\beta), \theta - \beta \rangle + \frac{1}{2\mu} \|\nabla g(\theta) - \nabla g(\beta)\|^2 \ .$$

FAST RATE FOR STRONGLY CONVEX FUNCTIONS

- Assume g is μ-strongly convex and L-smooth.

$$\theta_{k+1} = \theta_k - L^{-1} \nabla g(\theta_k).$$

- Function value rate:

$$g(\theta_k) - g(\theta_*) \leq \left(1 - \frac{\mu}{L}\right)^k [g(\theta_0) - g(\theta_*)]$$

FAST RATE FOR STRONGLY CONVEX FUNCTIONS

- Assume g is μ-strongly convex and L-smooth.

$$\theta_{k+1} = \theta_k - L^{-1}\nabla g(\theta_k).$$

- Function value rate:

$$g(\theta_k) - g(\theta_*) \leq \left(1 - \frac{\mu}{L}\right)^k [g(\theta_0) - g(\theta_*)]$$

- Applying the strongly convex property

$$g(\theta) \leq g(\eta) + \langle \nabla g(\eta), \theta - \eta \rangle + \frac{1}{2\mu}\|\nabla g(\theta) - \nabla g(\eta)\|^2 \ .$$

- Also implies linear-rate convergence for iterates:

$$\|\theta_k - \theta_*\|^2 \lesssim \mu^{-1}\left(1 - \frac{\mu}{L})\right)^k$$

SLOW RATE FOR SMOOTH CONVEX FUNCTIONS

- Assume g is convex and L-smooth, and let $\theta_* \in \arg\min_\theta g(\theta)$.
- Gradient Descent

$$\theta_{k+1} = \theta_k - L^{-1}\nabla g(\theta_k).$$

- We lose linear rate, but achieve

$$g(\theta_k) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{t+4} \ .$$

SLOW RATE FOR SMOOTH CONVEX FUNCTIONS

- Assume g is convex and L-smooth, and let $\theta_* \in \arg\min_\theta g(\theta)$.
- Gradient Descent

$$\theta_{k+1} = \theta_k - L^{-1}\nabla g(\theta_k).$$

- We lose linear rate, but achieve

$$g(\theta_k) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{t+4} \ .$$

- The algorithm thus adapts to problem "niceness".
- Is $O(t^{-1})$ optimal amongst first-order methods?

SHARPNESS OF FIRST-ORDER METHODS RATE

- *First-order method*: Any iterative algorithm that selects θ_k in

$$\theta_0 + \text{span}\{\nabla g(\theta_0), \ldots, \nabla g(\theta_{k-1})\}$$

- We consider the class of convex L-smooth functions with a global minimizer θ_*.

SHARPNESS OF FIRST-ORDER METHODS RATE

- *First-order method*: Any iterative algorithm that selects θ_k in

$$\theta_0 + \text{span}\{\nabla g(\theta_0), \ldots, \nabla g(\theta_{k-1})\}$$

- We consider the class of convex L-smooth functions with a global minimizer θ_*.

- **Theorem:** for every $k \leq (d-1)/2$ and every θ_0, there exist functions $g \in \mathcal{F}_L$ such that for any first-order method,

$$g(\theta_k) - g(\theta_*) \geq \frac{3L}{32} \frac{\|\theta_0 - \theta_*\|^2}{(t+1)^2} \quad .$$

SHARPNESS OF FIRST-ORDER METHODS RATE

- *First-order method*: Any iterative algorithm that selects θ_k in

$$\theta_0 + \text{span}\{\nabla g(\theta_0), \ldots, \nabla g(\theta_{k-1})\}$$

- We consider the class of convex L-smooth functions with a global minimizer θ_*.

- **Theorem:** for every $k \leq (d-1)/2$ and every θ_0, there exist functions $g \in \mathcal{F}_L$ such that for any first-order method,

$$g(\theta_k) - g(\theta_*) \geq \frac{3L}{32} \frac{\|\theta_0 - \theta_*\|^2}{(t+1)^2} \quad .$$

- Gradient Descent does not match this. Alternative?

PROOF SKETCH

- For each k, let

$$g_k(\theta) := \frac{L}{8}\left((\theta^1)^2 + \sum_{i \leq k-1} (\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1\right) \quad .$$

PROOF SKETCH

- For each k, let

$$g_k(\theta) := \frac{L}{8}\left((\theta^1)^2 + \sum_{i \le k-1}(\theta^i - \theta^{i+1})^2 + (\theta^k)^2 - 2\theta^1\right) \quad .$$

  - g_k is L-smooth.
  - Its minimiser spans the first k coordinates and has closed-form expression.
  - For any first-order method starting from zero, after k' iterations $\theta_{k'}$ is supported in first k' coordinates.
  - The optimum in that subspace also admits closed-form solution.
  - Given iteration k, consider $g = g_{2k+1}$ and bound $\dfrac{g(\theta_k) - g(\theta_*)}{\|\theta_0 - \theta_*\|^2}$

ACCELERATION

- Nesterov's scheme.
- Interpretation in the Quadratic case.

- Continuous-time Interpretation
- Bregman Lagrangian
- Euler Lagrange equations

ACCELERATED GRADIENT DESCENT (NESTEROV, 1983)

- We assume g is convex and L-smooth, with minimum attained at θ_*.
- Consider

$$\theta_k = \eta_{k-1} - L^{-1}\nabla g(\eta_{k-1})$$

$$\eta_k = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1})$$

ACCELERATED GRADIENT DESCENT (NESTEROV, 1983)

- We assume g is convex and L-smooth, with minimum attained at θ_*.
- Consider

$$\theta_k = \eta_{k-1} - L^{-1}\nabla g(\eta_{k-1})$$
$$\eta_k = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1})$$

- **Theorem (Nesterov'83):**

$$g(\theta_k) - g(\theta_*) \leq \frac{2L\|\theta_0 - \theta_*\|^2}{(t+1)^2} \ .$$

  - Optimal rate amongst first-order methods.
  - Proof similar to gradient-descent, but longer (see e.g. Bubeck'15).

ACCELERATED GRADIENT DESCENT, STRONGLY CONVEX

- Assume now g is μ-strongly convex and L-smooth.
- Consider

$$\theta_k = \eta_{k-1} - L^{-1}\nabla g(\eta_{k-1})$$

$$\eta_k = \theta_k + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(\theta_k - \theta_{k-1})$$

ACCELERATED GRADIENT DESCENT, STRONGLY CONVEX

- Assume now g is μ-strongly convex and L-smooth.
- Consider

$$\theta_k = \eta_{k-1} - L^{-1}\nabla g(\eta_{k-1})$$

$$\eta_k = \theta_k + \frac{1-\sqrt{\mu/L}}{1+\sqrt{\mu/L}}(\theta_k - \theta_{k-1})$$

- Bound becomes

$$g(\theta_k) - g(\theta_*) \leq L(1-\sqrt{\mu/L})^k \|\theta_0 - \theta_*\|^2$$

  - Better than with GD: $(1-\sqrt{\kappa})^k$ vs $(1-\kappa)^k$
  - Optimal!

INTERPRETATION OF ACCELERATED GRADIENT DESCENT

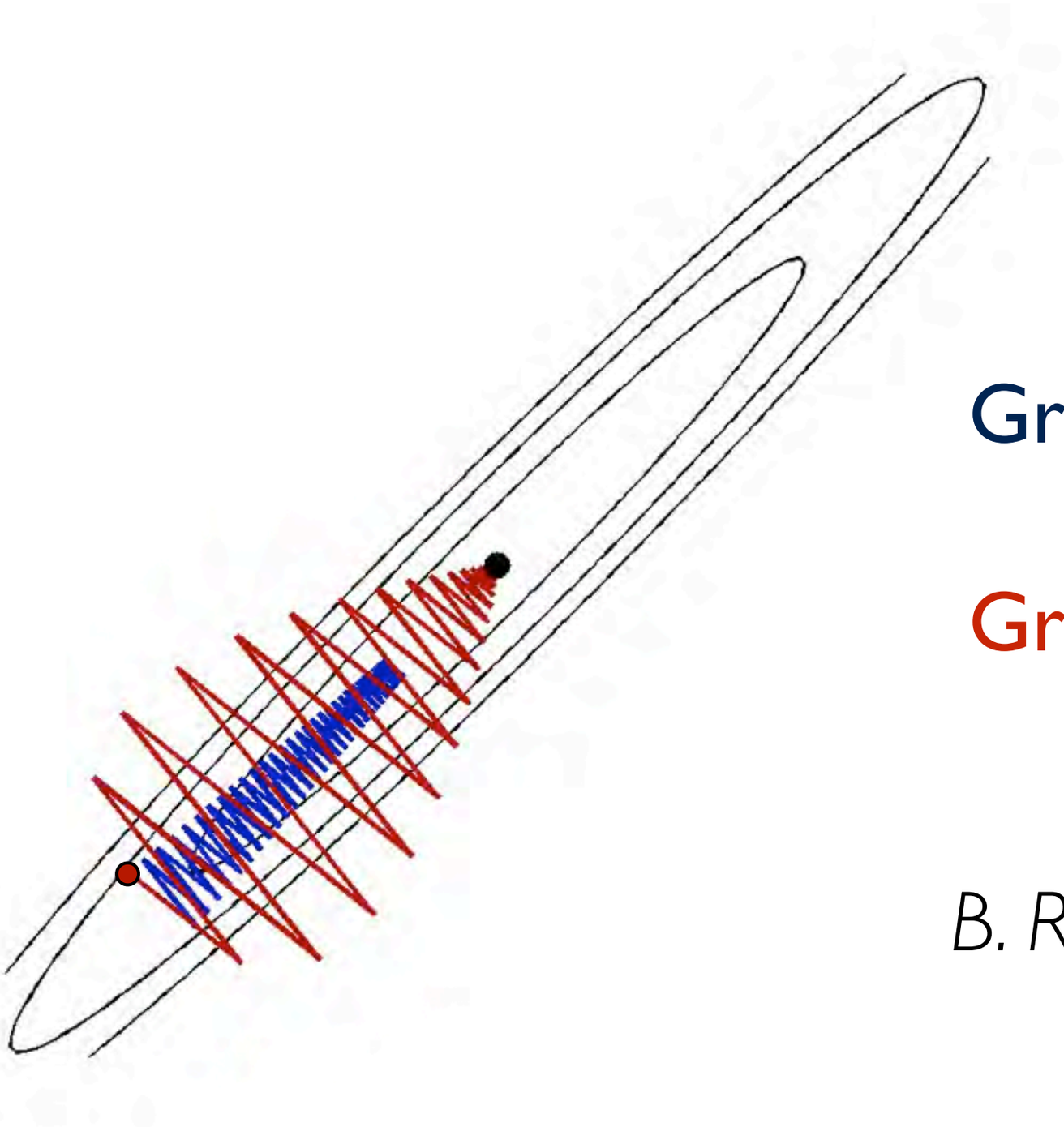
- Nesterov's method is typically associated to a *momentum* term:

$$\theta_k = \eta_{k-1} - L^{-1}\nabla g(\eta_{k-1})$$
$$\eta_k = (1-\gamma_k)\theta_k + \gamma_k\theta_{k-1} \ .$$

$$\eta_{k+1} - \eta_k = -\gamma_k[\eta_k - \eta_{k-1}] + \frac{\gamma_t}{L}[\nabla g(\eta_k) - \nabla g(\eta_{k-1})]$$

INTERPRETATION OF ACCELERATED GRADIENT DESCENT

- Q: Why does it work?



Gradient Descent

Accelerated Gradient Descent

figure credit: B. Recht [Simons'13]

INTERPRETATION OF ACCELERATED GRADIENT *(from M. Hardt)*

- Suppose we want to minimize

$$f(x) = \frac{1}{2} x^T A x - b^T x \ .$$

$A \in \mathbb{R}^{n \times n}$ positive definite.

INTERPRETATION OF ACCELERATED GRADIENT *(from M. Hardt)*

- Suppose we want to minimize

$$f(x) = \frac{1}{2} x^T A x - b^T x \ .$$

$A \in \mathbb{R}^{n \times n}$ positive definite.

- Its unique minimum is at $x^* = A^{-1} b$:

$$\nabla f(x) = 0 \Leftrightarrow Ax = b \ .$$

INTERPRETATION OF ACCELERATED GRADIENT *(from M. Hardt)*

- Suppose we want to minimize

$$f(x) = \frac{1}{2} x^T A x - b^T x \ .$$

$A \in \mathbb{R}^{n \times n}$ positive definite.

- Its unique minimum is at $x^* = A^{-1} b$:

$$\nabla f(x) = 0 \Leftrightarrow Ax = b \ .$$

- We run Gradient Descent starting at $x_0 = 0$ with step t:

$$x_{k+1} = x_k - t \nabla f(x_k) = [I - tA] x_k + tb \ .$$

- At iteration k we thus have

$$x_{k+1} = \left(\sum_{j \leq k} (I - tA)^j \right) (tb) \ .$$

INTERPRETATION OF ACCELERATED GRADIENT *(from M. Hardt)*

- Let $0 < l \leq L < \infty$ be the spectral bounds of A:

$$\forall\ x\ ,\ l\|x\| \leq \|Ax\| \leq L\|x\|\ .$$

INTERPRETATION OF ACCELERATED GRADIENT *(from M. Hardt)*

- Let $0 < l \leq L < \infty$ be the spectral bounds of A:

$$\forall\ x\ ,\ l\|x\| \leq \|Ax\| \leq L\|x\|\ .$$

- $\Rightarrow$ Eigenvalues of $(I - tA) \in (0, 1)$ if $t < L^{-1}$.
  - Thus $(tA)^{-1} = [I - (I - tA)]^{-1} = \sum_{j=0}^{\infty}(I - tA)^j$

$$\|(tA)^{-1} - \sum_{j=0}^{k}(I - tA)^j\| = O(\|(I - tA)^k\|) = O((1 - \frac{l}{L})^k)\ .$$

INTERPRETATION OF ACCELERATED GRADIENT *(from M. Hardt)*

- Let $0 < l \leq L < \infty$ be the spectral bounds of A:

$$\forall\ x\ ,\ l\|x\| \leq \|Ax\| \leq L\|x\|\ .$$

- $\Rightarrow$ Eigenvalues of $(I - tA) \in (0, 1)$ if $t < L^{-1}$.
- Thus $(tA)^{-1} = [I - (I - tA)]^{-1} = \sum_{j=0}^{\infty}(I - tA)^j$

$$\|(tA)^{-1} - \sum_{j=0}^{k}(I - tA)^j\| = O(\|(I - tA)^k\|) = O((1 - \frac{l}{L})^k)\ .$$

- This corresponds to the rate of standard gradient descent for strongly convex functions:

$$\|x_k - x^*\| \leq (1 - 2(\kappa + 1)^{-1})^k\|x_0 - x^*\|\ .$$

INTERPRETATION OF ACCELERATED GRADIENT DESCENT

➤We are thus approximating $(tA)^{-1}$ with a polynomial $q_k(A)$ of degree k.

INTERPRETATION OF ACCELERATED GRADIENT DESCENT

- We are thus approximating $(tA)^{-1}$ with a polynomial $q_k(A)$ of degree k.
- Q: What is the best polynomial approximation in our setting?

$$\text{for each } k \ , \ \min_{q_k} \|I - Aq_k(A)\| \ .$$

INTERPRETATION OF ACCELERATED GRADIENT DESCENT

- We are thus approximating $(tA)^{-1}$ with a polynomial $q_k(A)$ of degree k.
- Q: What is the best polynomial approximation in our setting?

$$\text{for each } k \ , \ \min_{q_k} \|I - Aq_k(A)\| \ .$$

A: Since A has eigenvalues in $[l, L]$, and we need $q_k(0) = 1$, Chebyshev polynomials are optimal:

INTERPRETATION OF ACCELERATED GRADIENT DESCENT

- We are thus approximating $(tA)^{-1}$ with a polynomial $q_k(A)$ of degree k.
- Q: What is the best polynomial approximation in our setting?

$$\text{for each } k \ , \ \min_{q_k} \|I - Aq_k(A)\| \ .$$

A: Since A has eigenvalues in $[l, L]$, and we need $q_k(0) = 1$, Chebyshev polynomials are optimal:

- **Lemma:** There is a polynomial p_k of degree $O(\sqrt{(L/l)\log(\epsilon^{-1})})$ such that $p_k(0) = 1$ and $|p_k(x)| \leq \epsilon$ for all $x \in [l, L]$.

Moreover, p_k can be computed recursively from previous **two** polynomials. It results that

$$x_{k+1} = x_k + \alpha_k \nabla f(x_k) + \beta_k \nabla f(x_{k-1}) \text{ for suitable } \alpha_k, \beta_k \ .$$

gives a convergence rate $\|x_k - x^*\| = O(\beta^k)$ with $\beta = 1 - 2(\sqrt{\kappa} + 1)^{-1}$

("minimax" optimal)

ACCELERATION

- Nesterov's scheme.
- Interpretation in the Quadratic case.
- Continuous-time Interpretation

- Bregman Lagrangian generalization
- Euler Lagrange equations
- Lyapunov

CONTINUOUS VS DISCRETE TIME

- Consider a gradient-descent scheme of the form

$$\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k)$$

- As $\gamma \to 0$, the trajectories of optimization $(\theta_k)_k$ converge to continuous curves.

CONTINUOUS VS DISCRETE TIME

- Consider a gradient-descent scheme of the form

$$\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k)$$

- As $\gamma \to 0$, the trajectories of optimization $(\theta_k)_k$ converge to continuous curves.

- These curves can be modeled as solutions of appropriate Ordinary Differential Equations (ODEs).

- In the case of gradient descent, the associated ODE is the *gradient flow:*

$$\dot{\theta} + \nabla g(\theta) = 0 \ .$$

$$\theta : [0, \infty) \to \mathbb{R}^d$$
$$\dot{\theta}(t) = \frac{\mathrm{d}\theta}{\mathrm{d}t}(t)$$

CONTINUOUS VS DISCRETE TIME

- Consider a gradient-descent scheme of the form

$$\theta_{k+1} = \theta_k - \gamma \nabla g(\theta_k)$$

- As $\gamma \to 0$, the trajectories of optimization $(\theta_k)_k$ converge to continuous curves.

- These curves can be modeled as solutions of appropriate Ordinary Differential Equations (ODEs).

- In the case of gradient descent, the associated ODE is the *gradient flow*:

$$\dot{\theta} + \nabla g(\theta) = 0 \ .$$

$$\theta : [0, \infty) \to \mathbb{R}^d$$
$$\dot{\theta}(t) = \frac{\mathrm{d}\theta}{\mathrm{d}t}(t)$$

- How to analyze Nesterov's scheme for general convex functions?

ORDINARY DIFFERENTIAL EQUATIONS

- We consider the Nesterov accelerated gradient scheme, assuming g is convex and L-smooth.

$$\theta_k = \eta_{k-1} - s\nabla g(\eta_{k-1})$$
$$\eta_k = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1}) \qquad 0 \leq s \leq L^{-1}$$

ORDINARY DIFFERENTIAL EQUATIONS

- We consider the Nesterov accelerated gradient scheme, assuming g is convex and L-smooth.

$$\theta_k = \eta_{k-1} - s\nabla g(\eta_{k-1})$$
$$\eta_k = \theta_k + \frac{k-1}{k+2}(\theta_k - \theta_{k-1}) \qquad 0 \leq s \leq L^{-1}$$

- Combining two equations and rescaling gives

$$\frac{\theta_{k+1} - \theta_k}{\sqrt{s}} = \frac{k-1}{k+2}\frac{\theta_k - \theta_{k-1}}{\sqrt{s}} - \sqrt{s}\nabla g(\eta_k)\,.$$

- Identify θ_k with $X(k\sqrt{(s)})$, for some smooth curve $X(t)$ $, t \geq 0$.

ODE AND NESTEROV

- Using the equivalence $k = t/\sqrt{s}$ and $X(t) \approx \theta_k$, we have also $X(t + \sqrt{s}) \approx \theta_{k+1}$.
- A Taylor expansion gives

$$\frac{\theta_{k+1} - \theta_k}{\sqrt{s}} = \dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) \ ,$$

$$\frac{\theta_k - \theta_{k-1}}{\sqrt{s}} = \dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) \ ,$$

$$\sqrt{s}\nabla g(\eta_k) = \sqrt{s}\nabla g(X(t)) + o(\sqrt{s}) \ .$$

ODE AND NESTEROV

➤ Substituting on the finite-difference equation, we obtain

$$\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) = \left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla g(X(t)) + o(\sqrt{s}) \ .$$

ODE AND NESTEROV

- Substituting on the finite-difference equation, we obtain

$$\dot{X}(t) + \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s}) = \left(1 - \frac{3\sqrt{s}}{t}\right)\left(\dot{X}(t) - \frac{1}{2}\ddot{X}(t)\sqrt{s} + o(\sqrt{s})\right) - \sqrt{s}\nabla g(X(t)) + o(\sqrt{s}) \ .$$

- The coefficients in $\sqrt{s}$ thus yield

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla g(X) = 0 \ .$$

- Initial conditions:
  - $X(0) = \theta_0 \ ,$
  - $\dot{X}(0) = \lim_{\sqrt{s}\to 0} \frac{\theta_2 - \theta_1}{\sqrt{s}} = -\lim_{\sqrt{s}\to 0} \sqrt{s}\nabla g(\eta_1) = 0 \ .$

ODE AND NESTEROV

- Denote $\mathcal{F}_L$: Class of L-smooth convex functions.
- **Theorem [Su et al.14]:** For any $g \in \mathcal{F}_L$ and any $\theta_0 \in \mathbb{R}^n$, the previous ODE with init condition $X(0) = \theta_0$, $\dot{X}(0) = 0$ has a unique global solution $X \in \mathcal{C}^2((0,\infty), \mathbb{R}^n) \cap \mathcal{C}^1([0,\infty), \mathbb{R}^n)$.
  - The ODE is well-posed despite being singular at $t = 0$.
- **Theorem [Su et al.14]:** For any $g \in \mathcal{F}_L$, as $s \to 0$, Nesterov scheme converges to the previous ODE: for all $T > 0$,

$$\lim_{s \to 0} \max_{k \leq T/\sqrt{s}} \|\theta_k - X(k\sqrt{s})\| = 0 \ .$$

  - Nesterov is a proper discretization.
  - Despite being a first-order method, ODE is 2nd order.

CONSEQUENCES: TIME AND ROTATIONAL INVARIANCE

- Consider a linear time dilation $\tilde{t} = ct,\ c > 0.$
- The reparametrised ODE becomes

$$\frac{\mathrm{d}^2 X}{\mathrm{d}\tilde{t}^2} + \frac{3}{\tilde{t}}\frac{\mathrm{d}X}{\mathrm{d}\tilde{t}} + \frac{\nabla g(X)}{c^2} = 0\ .$$

- Since minimizing $g, g/c^2$ is equivalent, model is invariant to time dilations.
- Coefficient in $1/t$ is necessary and sufficient.
- Similarly, we verify that the ODE is invariant to orthogonal transformations, as is the case with discrete gradient/ Nesterov.

CONSEQUENCES: ANALOGOUS CONVERGENCE RATE

- Recall that discrete Nesterov scheme with step s satisfies

$$g(\theta_k) - g(\theta_*) \leq \frac{2\|\theta_0 - \theta_*\|^2}{s(k+1)^2}$$

- The continuous ODE satisfies similar quadratic rate:

Theorem [Su et al'14]: For any $g \in \mathcal{F}_L$, let $X(t)$ be the unique solution of the ODE with initial conditions $X(0) = \theta_0$ and $\dot{X}(0) = 0$. Then

$$g(X(t)) - g(\theta_*) \leq \frac{2\|\theta_0 - \theta_*\|^2}{t^2} \ .$$

- Consistent with the sampling rate $t = k\sqrt{s}$.

NESTEROV VS GRADIENT DESCENT

- The equivalence $t \approx k\sqrt{s}$ suggests that running Nesterov's scheme for different learning rates $s, s' (\leq L^{-1})$ will produce iterates $(\theta_k)_k, (\theta'_{k'})_{k'}$ such that

$$\theta_k \approx \theta'_{k'} \text{ if } k\sqrt{s} \approx k'\sqrt{s'} \ .$$

- The integral curve that solves the ODE "contains" discrete iterates for different step sizes.
- Each Nesterov iteration travels $\sqrt{s}$ units of time.
- In contrast, each Gradient descent step moves s units of time along integral cuve of gradient flow $\dot{X} + \nabla g(X) = 0$.

CONSTANT 3?

- What is the role of the constant 3 in the ODE

$$\ddot{X} + \frac{3}{t}\dot{X} + \nabla g(X) = 0 \ .$$

- **Theorem [Su et al'14]:** If $r \geq 3$, the solution of $\ddot{X} + \frac{r}{t}\dot{X} + \nabla g(X) = 0$ satisfies

$$g(X(t)) - g(\theta^*) \leq \frac{(r-1)^2 \|\theta_0 - \theta_*\|^2}{2t^2} \ .$$

- If $r < 3$, inverse quadratic rate is lost: Phase transition.

FROM CONTINUOUS TO DISCRETE TIME

- The previous work derived a continuous-time ODE that is the limit of a given discrete optimization scheme.
- Can we take the opposite route?

- Given convex g, a general procedure to yield acceleration:

1. Construct and ODE such that its solutions $X(t)$ satisfy
$$g(X(t)) - g(\theta_*) \lesssim t^{-p}$$

2. Discretize the ODE in such a way to preserve convergence.

THE BREGMAN LAGRANGIAN [WIBISONO, WILSON, JORDAN]

- Assume the minimisation $\min_{\theta \in \Theta} g(\theta)$ admits a unique minimiser θ_* .
- Consider an auxiliary convex function h, to define a notion of distance in Θ via the *Bregman divergence*:

$$D_h(\theta, \eta) = h(\theta) - h(\eta) - \langle \nabla h(\eta), \theta - \eta \rangle$$

- non-negative thanks to convexity of h .
- locally equivalent to the Hessian metric

$$D_h(\theta, \eta) = \frac{1}{2}(y - x)^\top \nabla^2 h(x)(y - x) + o(\|y - x\|)$$

THE BREGMAN LAGRANGIAN

- The *Bregman Lagrangian* is defined as

$$\mathcal{L}(X, V, t) := e^{\alpha(t)+\gamma(t)} \left(D_h(X + e^{-\alpha(t)} V, X) - e^{\beta(t)} g(X) \right) .$$

X: position, V: velocity, t: time.

- The functions $\alpha(t), \gamma(t), \beta(t)$ satisfy *ideal scaling* if

$$\dot{\beta}(t) \leq e^{\alpha(t)} \ , \ \dot{\gamma}(t) = e^{\alpha(t)}.$$

- Given a path in space-time $X_t \in \Theta; t \geq 0$, the *action* is obtained by integrating the Lagrangian of the system:

$$\mathcal{J}(X) = \int_{t \geq 0} \mathcal{L}(X_t, \dot{X}_t, t) dt \, .$$

THE BREGMAN LAGRANGIAN

- The minimal action curves necessarily satisfy the *Euler-Lagrange equation*:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{\partial \mathcal{L}}{\partial V} = \frac{\partial \mathcal{L}}{\partial X}\,.$$

- In the case of the Bregman Lagrangian with ideal scaling, this equation becomes

$$\frac{\mathrm{d}}{\mathrm{d}t}\nabla h(X_t + e^{-\alpha(t)}\dot{X}_t) = -e^{\alpha(t)+\beta(t)}\nabla g(X_t)\,.$$

- Do the solutions minimize g ? How fast?

THE BREGMAN LAGRANGIAN

- **Theorem [Wibisono et al'17]:** Under ideal scaling, the solutions to the Euler-Lagrange equation satisfy

$$g(X_t) - g(\theta_*) \leq O(e^{-\beta(t)}) \,.$$

- Consider the Lyapunov function

$$\mathcal{E}_t = D_h(\theta_*, X_t + e^{-\alpha(t)}\dot{X}_t) + e^{\beta(t)}\left(g(X_t) - g(\theta_*)\right) \,.$$

- For a given $\alpha(t)$, the optimal convergence rate is achieved by $\dot{\beta}(t) = e^{\alpha(t)}$.

- We want fast convergence rate, but such that discretization preserves the rate!

THE BREGMAN LAGRANGIAN

- Consider the Bregman Lagrangians generated by parameters

$$\alpha(t) = \log p - \log t\,,\ \beta(t) = p\log t + \log C\,,\ \gamma(t) = p\log t\,.$$
$$(p > 0, C > 0)\ .$$

- They satisfy the ideal scaling condition, and the resulting Euler-Lagrange equation is

$$\ddot{X}_t + \frac{p+1}{t}\dot{X}_t + Cp^2 t^{p-2}\left[\nabla^2 h(X_t + tp^{-1}\dot{X}_t)\right]^{-1}\nabla g(X_t) = 0\,.$$

  - From previous theorem, convergence rate $O(t^{-p})$.
  - $p = 2$ and $h(x) = \frac{1}{2}\|x\|^2$ is the ODE from Su et al.

DISCRETIZING EULER LAGRANGE EQUATIONS

- The decoupled first-order system of equations is

$$Z_t = X_t + \frac{t}{p}\dot{X}_t \,,\ \frac{d}{dt}\nabla h(Z_t) = -Cpt^{p-1}\nabla g(X_t) \ .$$

- Discrete Euler scheme becomes

$$\eta_k = \arg\min_z \left\{ Cpk^{p-1}\langle \nabla g(\theta_k), z\rangle + \delta^{-p} D_h(z, \eta_{k-1}) \right\}$$

$$\theta_{k+1} = \frac{p}{k}\eta_k + \frac{k-p}{k}\theta_k$$

- It turns out that the simple forward-backward Euler method to discretize this ODE is not stable, thus does not produce a discrete optimization scheme with matching convergence rate.

DISCRETIZING EULER-LAGRANGE EQUATIONS

- Nesterov's constructions on non-Euclidean domains (mirror descent) provide a general stable discretization scheme with matching rates.

$$\theta_{k+1} = \frac{p}{k+p}\eta_k + \frac{k}{k+p}\xi_k$$

$$\eta_k = \arg\min_z \left\{ Cpk^{p-1}\langle \nabla g(\xi_k), z\rangle + \delta^{-p} D_h(z, \eta_{k-1}) \right\}$$

with ξ_k satisfying

$$\langle \nabla g(\xi_k), \theta_k - \xi_k\rangle \geq M\delta^{p/(p-1)} \|\nabla g(\xi_k)\|^{p/(p-1)} .$$

- generalization of co-coercivity property.
- explicit construction for high-order accelerated gradients.

FROM CONTINUOUS TO DISCRETE TIME

- Other related works:
  - Lyapunov Analysis
  - Symplectic Optimization.

- Extensions: composite optimization, non-Euclidean settings

STOCHASTIC GRADIENT DESCENT

- sgd main properties
- convergence with fixed lr (bach moulines)
- sgd sde papers
  - two examples.