JHU vision lab

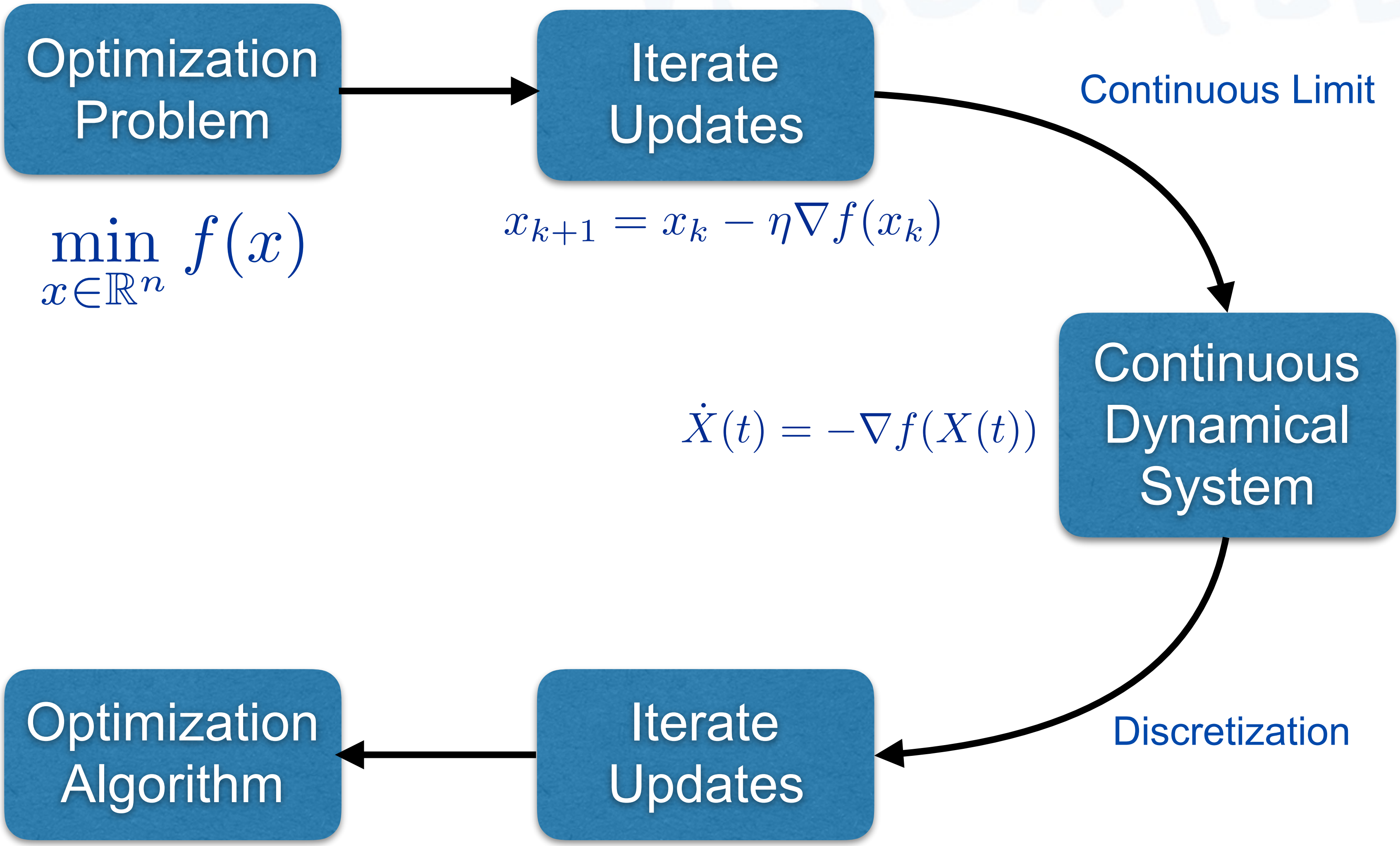
# Dynamical Systems and ADMM

**René Vidal**
Herschel Seder Professor of Biomedical Engineering
Director of the Mathematical Institute for Data Science
Johns Hopkins University

THE DEPARTMENT OF BIOMEDICAL ENGINEERING
The Whitaker Institute at Johns Hopkins

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
*for* DATA SCIENCE

# Optimization and Dynamical Systems




[1] Cauchy, 1847
[2] Su, Boyd, Candes, NIPS 2014, JMLR 2016
[3] Wibisono, Wilson, Jordan, PNAS 2016
[4] Attouch, Chbani, Peypouquet, Redont, Math. Prog. 2016
[5] Krichene, Bayen, Bartlett, NIPS 2015
[6] Fazlyab, Ribeiro, Morari, Preciado, 2017
[7] Fazlyab, et al. SIAM Opt 2018
[8] Lessard, Recht, Packard, SIAM Opt 2016

# Gradient Flow (simplest example)

- Unconstrained optimization problem $$\min_{x \in \mathbb{R}^n} f(x)$$
- Gradient descent (GD) $$x_{k+1} = x_k - \eta \nabla f(x_k)$$
- Discretization of **Gradient Flow** (GF) [1] $$\dot{X}(t) = -\nabla f(X(t))$$
- Convergence rate for **convex functions**

| | $f(x) - f(x^\star)$ |
|---|---|
| Gradient Descent | $O(1/k)$ |
| Gradient Flow | $O(1/t)$ |

[1] Cauchy, Augustin (1847). Méthode générale pour la résolution des systèmes d'équations simultanées. Comptes Rendus Hebd. Séances Acad. Sci. 25:536–538.

# Accelerated Gradient Flow

- Nesterov's **Accelerated Gradient Descent** (AGD) [1]

$$x_{k+1} = \hat{x}_k - \eta \nabla f(\hat{x}_k)$$

$$\hat{x}_{k+1} = x_{k+1} + \frac{k}{k+r}(x_{k+1} - x_k)$$

- Discretization of the **Accelerated Gradient Flow** (AGF) [2]

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

- Convergence rate for convex functions (optimal rate)

| | $f(x) - f(x^\star)$ |
|---|---|
| Accelerated Gradient Descent | $O(1/k^2)$ |
| Accelerated Gradient Flow | $O(1/t^2)$ |

[1] Nesterov. A Method of Solving a Convex Programming Problem with Convergence Rate O(1/k²).
Soviet Mathematics Doklady, 27(2):372–376, 1983.
[2] Su, Boyd, Candes, NIPS 2014, JMLR 2016

# Our Contributions

- **Prior work**
  - Smooth functions, unconstrained problems, gradient based
  - Many problems in machine learning are constrained and non-smooth, e.g. sparse regularization (L1 norm), nuclear norm minimization, etc.

- **Our work**
  - Accelerated, proximal based, linearly constrained [1]; non-smooth [2]

$$\min_{x,z}\{\Phi(x,z) \equiv f(x) + g(z)\} \quad \text{s.t.} \quad z = Ax$$

- **Contributions**
  - Differential equations/inclusions for variants of (accelerated) ADMM
  - Lyapunov stability
  - Convergence rates
  - New variants including relaxation + acceleration

[1] França, Robinson, Vidal, ICML 2018
[2] França, Robinson, Vidal, arXiv: 1805.06579 2018

# ADMM Flow

$$\min_{x,z} \underbrace{f(x) + g(z)}_{\Phi(x,z)} + \langle \rho u, Ax - z \rangle + \frac{\rho}{2}\|Ax - z\|^2$$

- Alternating Direction Method of Multipliers (ADMM) [1,2]

$$x_{k+1} = \operatorname{argmin}_x \; f(x) + (\rho/2)\|Ax - z_k + u_k\|^2$$

$$z_{k+1} = \operatorname{argmin}_z \; g(z) + (\rho/2)\|Ax_{k+1} - z + u_k\|^2$$

$$u_{k+1} = u_k + Ax_{k+1} - z_{k+1}$$

- ADMM is popular in large scale applications of machine learning and statistics (easily distributed) [3]

**Theorem** [4] *The continuous limit of ADMM is the ADMM Flow*

$$(A^T A)\dot{X}(t) = -\nabla \Phi(X(t))$$


[1] Gabay, Mercier, Comp. Math. App., 1976
[2] Glowinsky, Marroco, 1975
[3] Boyd, Parikh, Chu, Peleato, Eckestein, Found. Trends in Mach. Learning, 2011
[4] França, Robinson, Vidal, ICML 2018

# Accelerated ADMM Flow

- Accelerated ADMM (A-ADMM) [1]

$$
\begin{aligned}
x_{k+1} &= \operatorname{argmin}_x \; f(x) + (\rho/2)\|Ax - \hat{z}_k + \hat{u}_k\|^2 \\
z_{k+1} &= \operatorname{argmin}_z \; g(z) + (\rho/2)\|Ax_{k+1} - z + \hat{u}_k\|^2 \\
u_{k+1} &= \hat{u}_k + Ax_{k+1} - z_{k+1} \\
\hat{u}_{k+1} &= u_{k+1} + \tfrac{k}{k+r}(u_{k+1} - u_k) \\
\hat{z}_{k+1} &= u_{k+1} + \tfrac{k}{k+r}(u_{k+1} - u_k)
\end{aligned}
\qquad r \geq 3
$$

**Theorem** [2] *The continuous limit of A-ADMM is the A-ADMM Flow*

$$
(A^T A)\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) = -\nabla\Phi(X(t))
$$

- Generalizes previous results (linear constraint) [3,4]
- For now we assume differentiability (will be relaxed later)


[1] Goldstein, O'Donoghue, Setzer, Baraniuk, SIAM Im. Sci., 2014
[2] França, Robinson, Vidal, ICML 2018
[3] Su, Boyd, Candes NIPS 2014, JMLR 2016
[4] Wibisono, Wilson, Jordan PNAS 2016

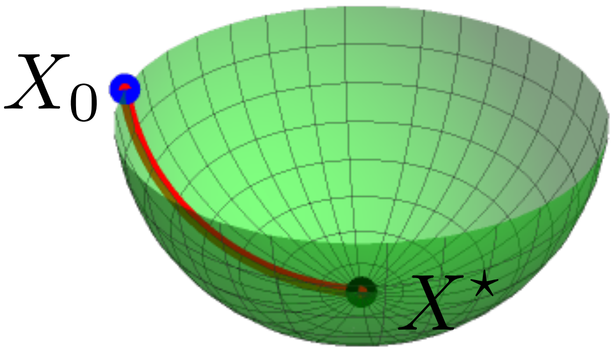
# Stability of ADMM Flow

- **Objective**: $\Phi(X) = f(X) + g(AX)$
  - f, g: cont. diff. & convex
  - A: full column rank

- **ADMM Flow**

$$(A^T A)\dot{X}(t) = -\nabla\Phi(X(t))$$

**Theorem [1]:** *Let $X^\star$ be a strict local minimizer and isolated critical point of $\Phi$. Then, $X^\star$ is* ***asymptotically stable****.*




**Theorem [1]:** *Let $\Phi$ be convex. Then,*

$$\Phi(X(t)) - \Phi(X^\star) \leq \frac{C}{t}$$

- Matches the known rate of ADMM [2] $O(1/k)$
- Proof based on Lyapunov functions


[1] França, Robinson, Vidal, ICML 2018
[2] Eckstein, J. and Yao, W. Understanding the Convergence of the Alternating Direction Method of Multipliers: Theoretical and Computational Perspectives. 2015.

# Stability of Accelerated ADMM Flow

- Objective function

$$\Phi(X) = f(X) + g(AX)$$

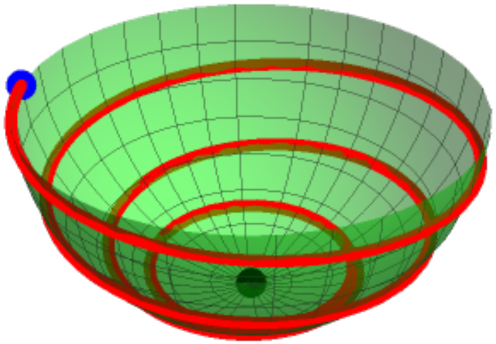
- A-ADMM Flow

$$(A^T A)\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) = -\nabla\Phi(X(t))$$

**Theorem.** *Let* $X^\star$ *be a strict local minimizer and isolated critical point of* $\Phi$*. Then,* $(X, \dot{X}) = (X^\star, 0)$ *is* ***stable****.*

**Theorem.** *Let* $\Phi$ *be convex and* $r \geq 3$*. Then,*

$$\Phi(X(t)) - \Phi(X^\star) \leq \frac{C}{t^2}$$



- Convergence rate currently unknown in discrete case, but thus suggests O(1/k²).

[1] França, Robinson, Vidal, ICML 2018

# Stability of Accelerated ADMM Flow

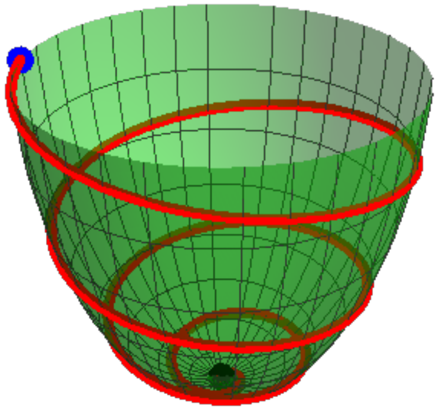
- **Definition:** Forcing function $\phi : [0,\infty) \to [0,\infty)$

$$\lim_{k\to\infty} \phi(\xi_k) = 0 \implies \lim_{k\to\infty} \xi_k = 0 \quad \forall\{\xi_k\}$$

**Theorem [1]** *Let* $X^\star$ *be a strict local minimizer of* $\Phi$ *such that for all* $X \in \mathcal{B}(X^\star)$ *we have* $\Phi(X) - \Phi(X^\star) \geq \phi(\|X - X^\star\|)$

Then, $(X, \dot{X}) = (X^\star, 0)$ is **asymptotically stable**.

- **Example:** uniformly convex functions



[1] França, Robinson, Vidal, ICML 2018

# Basic Proof Technique (Lyapunov functions)

- Stability follows from standard Lyapunov's theorem

$$\mathcal{E}(X) > 0 \quad \text{and} \quad \dot{\mathcal{E}} \le 0 \quad \Longrightarrow \quad \textbf{Stability}$$

$$\mathcal{E}(X) > 0 \quad \text{and} \quad \dot{\mathcal{E}} < 0 \quad \Longrightarrow \quad \textbf{Asymptotic Stability}$$

- Convergence rates ("basic idea")

$$\mathcal{E}(X, \dot{X}, t) = a(t)\left(\Phi(X) - \Phi(X^\star)\right) + \underbrace{\cdots}_{\ge 0} \ge 0$$

$$\dot{\mathcal{E}}(X, \dot{X}, t) \le 0$$

$$\Longrightarrow \quad \Phi(X(t)) - \Phi(X^\star) \le \frac{\mathcal{E}|_{t=0}}{a(t)} = \frac{C}{a(t)}$$

- Difficulty lies in **constructing** appropriate Lyapunov functions for each ODE, under convex, strongly convex, etc.

[1] França, Robinson, Vidal, ICML 2018
[2] França, Robinson, Vidal, arXiv: 1805.06579 2018

# Generalization to Non-smooth Problems

- So far we assumed the objective is differentiable
- This can be relaxed and we obtain differential inclusions

$$(A^T A)\dot{X}(t) \in -\partial\Phi(X(t)) \qquad \textbf{ADMM}$$

$$(A^T A)\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) \in -\partial\Phi(X(t)) \qquad \textbf{A-ADMM}$$

- The previous Lyapunov analysis can be generalized to non-smooth problems (directional derivatives, etc.). For instance,

$$\frac{d}{dt}(\Phi \circ X)(t) = \langle g, \dot{X}(t)\rangle \quad \forall g \in \partial\Phi(X(t)) \quad (a.e)$$

- Previous rates and stability remains the same

[1] Aubin, Cellina, Differential Inclusions (1984)
[2] Clarke, Nonsmooth Analysis and Control Theory (2013)

# New Variants of Accelerated ADMM

- We introduce relaxation and two types of acceleration

$$x_{k+1} = \operatorname{argmin}_x \; f(x) + (\rho/2)\|Ax - \hat{z}_k + \hat{u}_k\|^2$$

$$z_{k+1} = \operatorname{argmin}_z \; g(z) + (\rho/2)\|\alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z + \hat{u}_k\|^2$$

$$u_{k+1} = \hat{u}_k + \alpha A x_{k+1} + (1-\alpha)\hat{z}_k - z_{k+1}$$

$$\hat{u}_{k+1} = u_{k+1} + \gamma_{k+1}(u_{k+1} - u_k)$$

$$\hat{z}_{k+1} = u_{k+1} + \gamma_{k+1}(u_{k+1} - u_k)$$

$$\gamma_{k+1} = \begin{cases} k/(k+r) & \text{Nesterov} \\ 1 - r/\sqrt{\rho} & \text{Heavy Ball} \end{cases}$$

$$r \geq 3$$

- Relaxation parameter $\alpha \in (0, 2)$
- Non relaxed version recovered with $\alpha = 1$
- Non accelerated (but relaxed) ADMM: $\gamma_{k+1} = 0$

We thus propose relaxed accelerated ADMM (**R-A-ADMM**) and relaxed Heavy Ball ADMM (**R-HB-ADMM**)

[1] França, Robinson, Vidal, arXiv: 1805.06579 2018

# Continuous Limit of Variants of ADMM

**Theorem.** In the continuous limit we obtain

$$(2-\alpha)(A^TA)\dot{X}(t) \in -\partial\Phi(X(t)) \quad \text{R-ADMM}$$

$$(2-\alpha)(A^TA)\left(\ddot{X}(t) + \frac{r}{t}\dot{X}(t)\right) \in -\partial\Phi(X(t)) \quad \text{R-A-ADMM}$$

$$(2-\alpha)(A^TA)\left(\ddot{X}(t) + r\dot{X}(t)\right) \in -\partial\Phi(X(t)) \quad \text{R-HB-ADMM}$$

- Generalizes [1,2] for non smooth and linear constraints
- The above are non smooth dynamical systems
- When $\Phi$ is convex lower semicontinuous existence of solutions is guaranteed

[1] França, Robinson, Vidal, arXiv: 1805.06579 2018
[2] Su, Boyd, Candes, NIPS 2014, JMLR 2016
[3] Wibisono, Wilson, Jordan, PNAS 2016

# Non-smooth Lyapunov Analysis

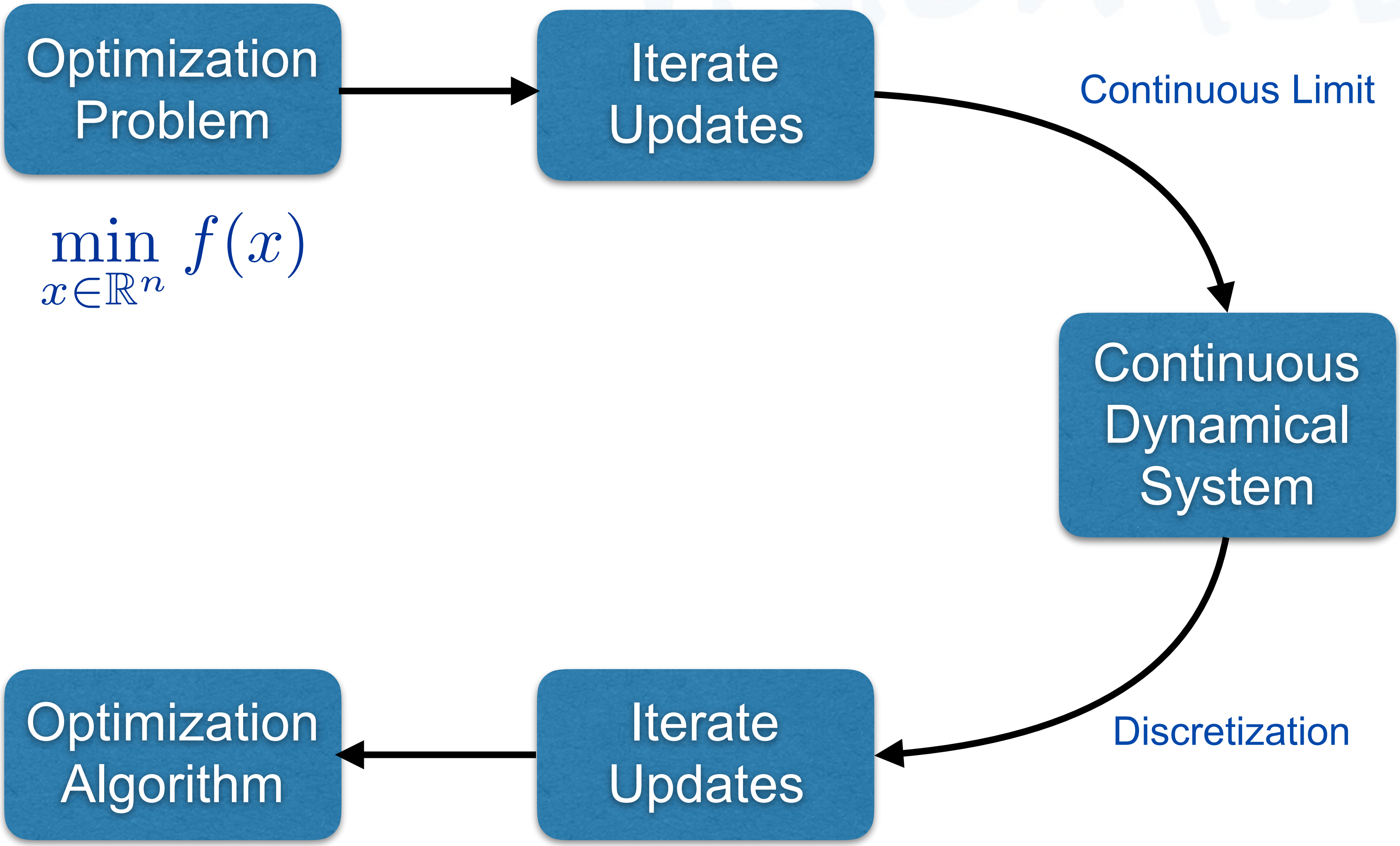
- We obtain the following rates in the continuous case

| | *convex* | *strongly convex* |
|---|---|---|
| ADMM | $O\left(\frac{\sigma_1^2(A)}{t}\right)$ | $O\left(\kappa(A)e^{-\mu t/(2\sigma_1^2(A))}\right)$ |
| A-ADMM† | $O\left(\frac{(r-1)^2\sigma_1^2(A)}{t^2}\right)$ | $O\left(\frac{(r\sigma_1(A))^{2r/3}}{\mu^{r/3}}\frac{1}{t^{2r/3}}\right)$ |
| R-ADMM† | $O\left(\frac{(2-\alpha)(\sigma_1^2(A))}{t}\right)$ | $O\left(\kappa(A)e^{-\mu t/(2(2-\alpha)\sigma_1^2(A))}\right)$ |
| R-A-ADMM‡ | $O\left(\frac{(2-\alpha)(r-1)^2\sigma_1^2(A)}{t^2}\right)$ | $O\left(\frac{((2-\alpha)^{1/2}r\sigma_1(A))^{2r/3}}{\mu^{r/3}}\frac{1}{t^{2r/3}}\right)$ |
| R-HB-ADMM‡ | $O\left(\frac{r(2-\alpha)\sigma_1^2(A)}{t}\right)$ | $O\left((2-\alpha)r^2\sigma_1^2(A)e^{-2rt/3}\right)$ |

- Most of these rates are unknown in the discrete case
- Interesting tradeoff between **Nesterov** vs. **Heavy Ball** acceleration in **convex** vs. **strongly convex** settings

[1] França, Robinson, Vidal, arXiv: 1805.06579 2018

# From Dynamical Systems to Optimization




$$\min_{x \in \mathbb{R}^n} f(x)$$

[1] Cauchy, 1847
[2] Su, Boyd, Candes, NIPS 2014, JMLR 2016
[3] Wibisono, Wilson, Jordan, PNAS 2016
[4] Attouch, Chbani, Peypouquet, Redont, Math. Prog. 2016
[5] Krichene, Bayen, Bartlett, NIPS 2015
[6] Fazlyab, Ribeiro, Morari, Preciado, 2017
[7] Fazlyab, et al. SIAM Opt 2018
[8] Lessard, Recht, Packard, SIAM Opt 2016

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
for DATA SCIENCE

# Hamiltonian Systems

- Accelerated non-smooth systems can be represented by non-smooth Hamiltonian systems [1,2,3]

$$H = \frac{1}{2}\langle P, M^{-1}P\rangle + \lambda\Phi(X)$$

$M = A^T A$ "mass matrix"

$\lambda = 1/(2-\alpha)$ "coupling constant"

Nonlinear harmonic oscillator

- We use a **Conformal Hamiltonian** formulation [4]

- Phase space volumes dissipate exponentially (damping)

- Previous differential inclusions are equivalent to

$$\dot{X} = \nabla_P H = M^{-1}P$$

$$\dot{P} \in -\partial_X H - \underbrace{\eta P}_{\text{dissipation}} = -\lambda\partial\Phi(X) - \eta P$$

$$\eta = \begin{cases} r/t & \text{Nesterov (not conformal)} \\ r & \text{Heavy Ball (conformal)} \end{cases}$$


[1] Rockafellar, 1970
[2] Clarke, 1976
[3] Ioffe, 1997
[4] McLachlan, Perlmutter, 2011

# Symplectic Integration

- Symplectic integration are discretization techniques that preserve symplectic structure of Hamiltonian systems
- Widely used in molecular dynamics simulations, statistical mechanics, Monte Carlo method, particle physics, etc.
- We use Strang splitting (leapfrog, Stormer-Verlett)

$$p_{k+1/2} \in p_k - (h/2)\lambda\partial\Phi(x_k)$$
$$x_{k+1} = x_k + hM^{-1}p_{k+1/2}$$
$$p_{k+1} \in e^{-\Delta_k}p_{k+1/2} - (h/2)\lambda\partial\Phi(x_k)$$

$$\Delta_k = \int_{t_k}^{t_{k+1}} \eta(t)dt \quad (dissipation)$$

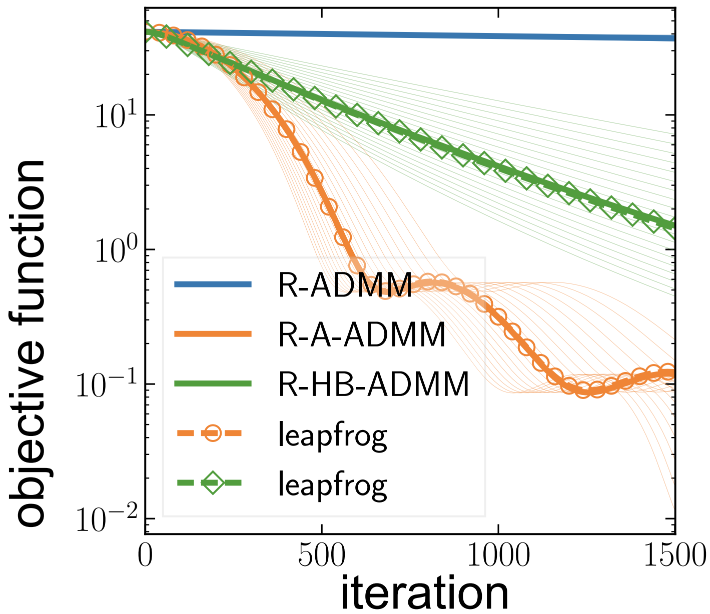
- This method has the same cost of (sub)gradient descent, one (sub)gradient per iteration

# Numerical Simulations

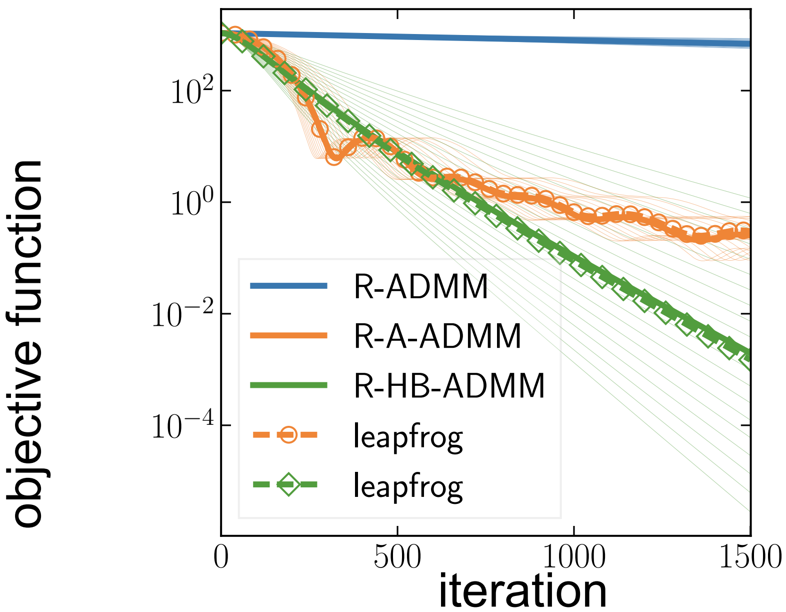
- Compare ADMM variants with dynamical simulations

$$\min_{x} \frac{1}{2} x^T Q x \quad \text{s.t.} \quad z = Ax$$

**Convex**




**Strongly Convex**



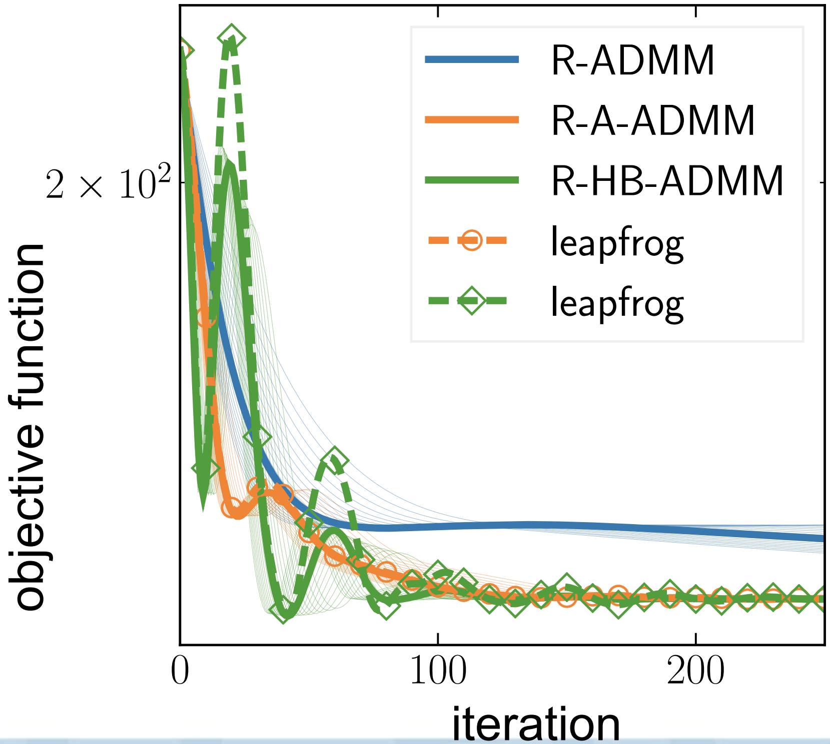

- Note the tradeoff: Nesterov vs. Heavy Ball (as predicted)
- Note agreement between algorithms and dynamical systems

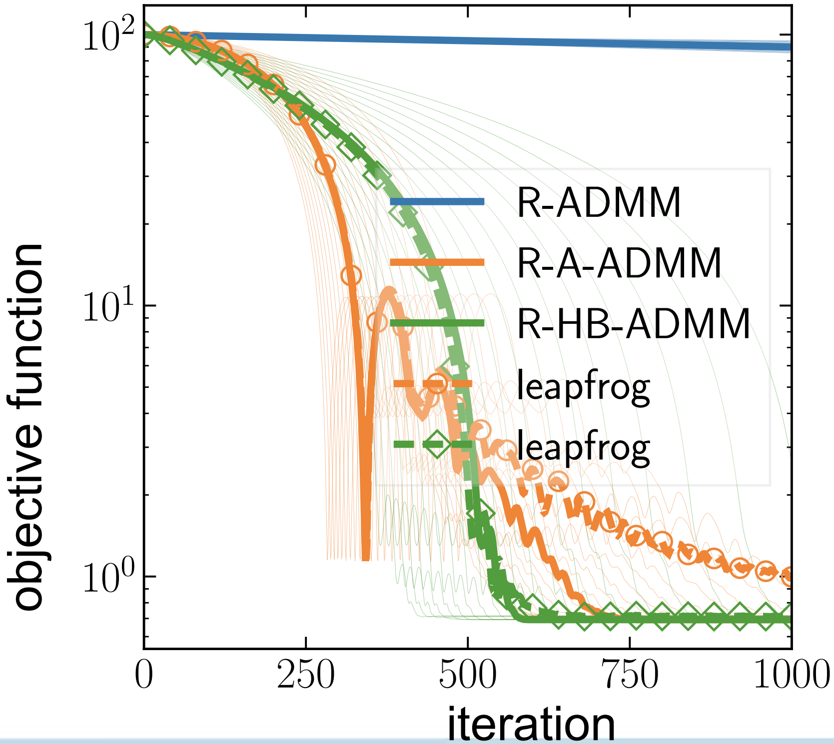
# Numerical Simulations

- Linear regression problem with elastic net regularization
- Sparse logistic regression

$$\min_x \tfrac{1}{2}\|y - Mx\|^2 + \|x\|_1 + \tfrac{1}{2}\|x\|_2^2$$

$$\min_w \sum_{i=1}^{n} \log(1 + e^{-y_i w^T x_i}) + \|w\|_1$$



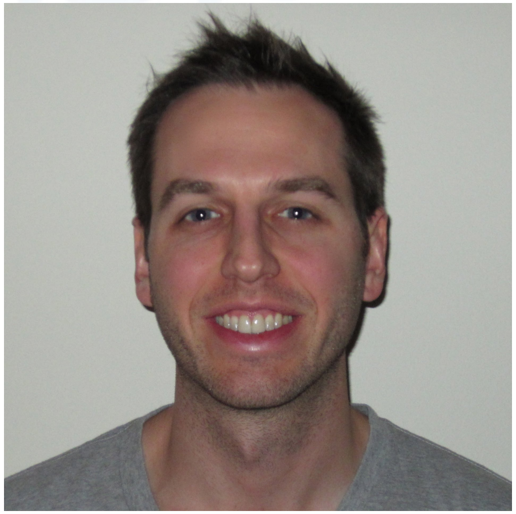
# Conclusions

- We considered known instances of (accelerated) ADMM, and proposed **new variants of accelerated ADMM**
- Connections with **non smooth dynamical systems**
- Generalize previous approaches (smooth, unconstrained, gradient based)
- Analyzed **stability**
- **New rate-of-convergence results** (continuous time)
- Conformal Hamiltonian description
- **Numerical simulations** of continuous dynamical systems in comparison to the algorithms (symplectic integration)

# Acknowledgements

Guilherme Franca

Daniel Robinson



Work supported by the Army Research Office under contract ARO MURI W911NF-17-1-0304 as part of the collaboration between US DOD, UK MOD and UK EPSRC under the Multidisciplinary University Research Initiative.

Mathematical Institute for Data Science @ JHU
http://www.minds.jhu.edu

JOHNS HOPKINS
MATHEMATICAL INSTITUTE
*for* DATA SCIENCE