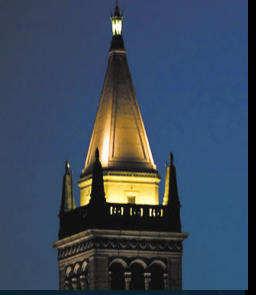


Optimization and Dynamical Systems: Variational, Hamiltonian, and Symplectic Perspectives

Michael Jordan

University of California, Berkeley

Computation and Statistics

- A Grand Challenge of our era: tradeoffs between statistical inference and computation
  - most data analysis problems have a time budget
  - and often they're embedded in a control problem
- Optimization has provided the computational model for this effort (computer science, not so much)
  - it's provided the algorithms and the insight
- On the other hand, modern large-scale statistics has posed new challenges for optimization
  - millions of variables, millions of terms, sampling issues, nonconvexity, need for confidence intervals, parallel/distributed platforms, etc

Computation and Statistics (cont)

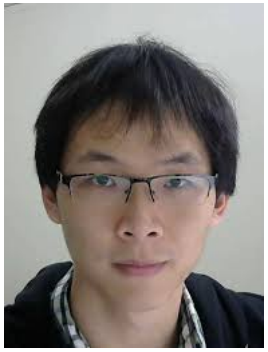
- Modern large-scale statistics has posed new challenges for optimization
  - millions of variables, millions of terms, sampling issues, nonconvexity, need for confidence intervals, parallel/distributed platforms, etc
- Current algorithmic focus: what can we do with the following ingredients?
  - gradients
  - stochastics
  - acceleration
- Current theoretical focus: placing lower bounds from statistics and optimization in contact with each other

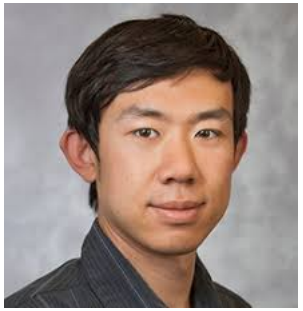
Outline

- Escaping saddle points efficiently
- Variational, Hamiltonian and symplectic perspectives on Nesterov acceleration
- Acceleration and saddle points
- Acceleration and Langevin diffusions
- Optimization and empirical processes

Part I: How to Escape Saddle Points Efficiently

with Chi Jin, Praneeth Netrapalli, Rong Ge, and Sham Kakade

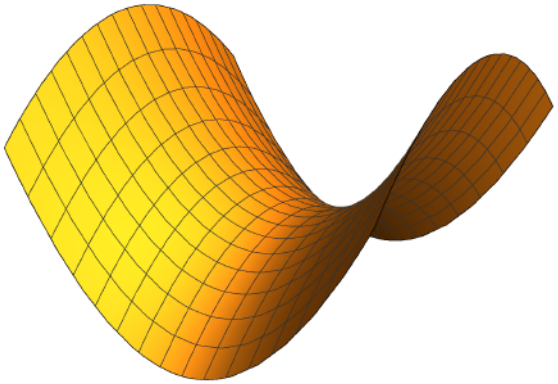




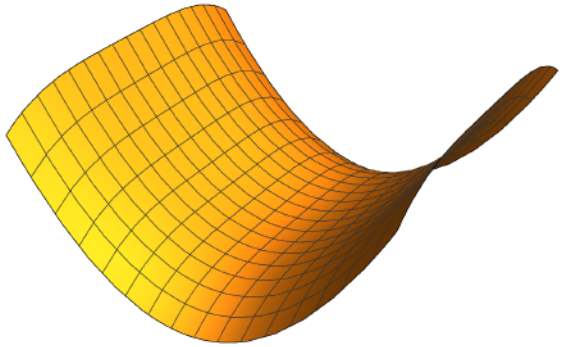

Nonconvex Optimization and Statisitics

- Many interesting statistical models yield nonconvex optimization problems (cf neural networks)
- Bad local minima used to be thought of as the main problem in fitting such models
- But in many convex problems there either are no local optima (provably), or stochastic gradient seems to have no trouble (eventually) finding global optima
- But saddle points abound in these architectures, and they cause the learning curve to flatten out, perhaps (nearly) indefinitely

The Importance of Saddle Points



Strict saddle point



Non-strict saddle point

- How to escape?
  - need to have a negative eigenvalue that's strictly negative
- How to escape efficiently?
  - in high dimensions how do we find the direction of escape?
  - should we expect exponential complexity in dimension?

A Few Facts

- Gradient descent will asymptotically avoid saddle points (Lee, Simchowitz, Jordan & Recht, 2017)
- Gradient descent can take exponential time to escape saddle points (Du, Jin, Lee, Jordan, & Singh, 2017)
- Stochastic gradient descent can escape saddle points in polynomial time (Ge, Huang, Jin & Yuan, 2015)
  - but that's still not an explanation for its practical success
- Can we prove a stronger theorem?

Optimization

Consider problem:

$$\min_{\mathbf{x} \in \mathbb{R}^d} \; f(\mathbf{x})$$

Gradient Descent (GD):

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t).$$

Optimization

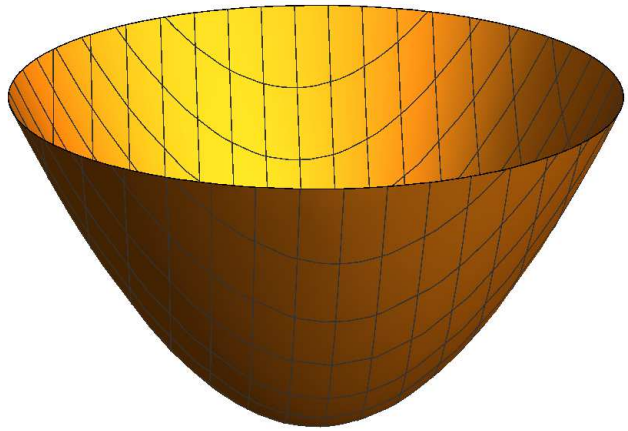
Consider problem:

$$\min_{\mathbf{x}\in\mathbb{R}^d} \; f(\mathbf{x})$$

Gradient Descent (GD):

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t).$$

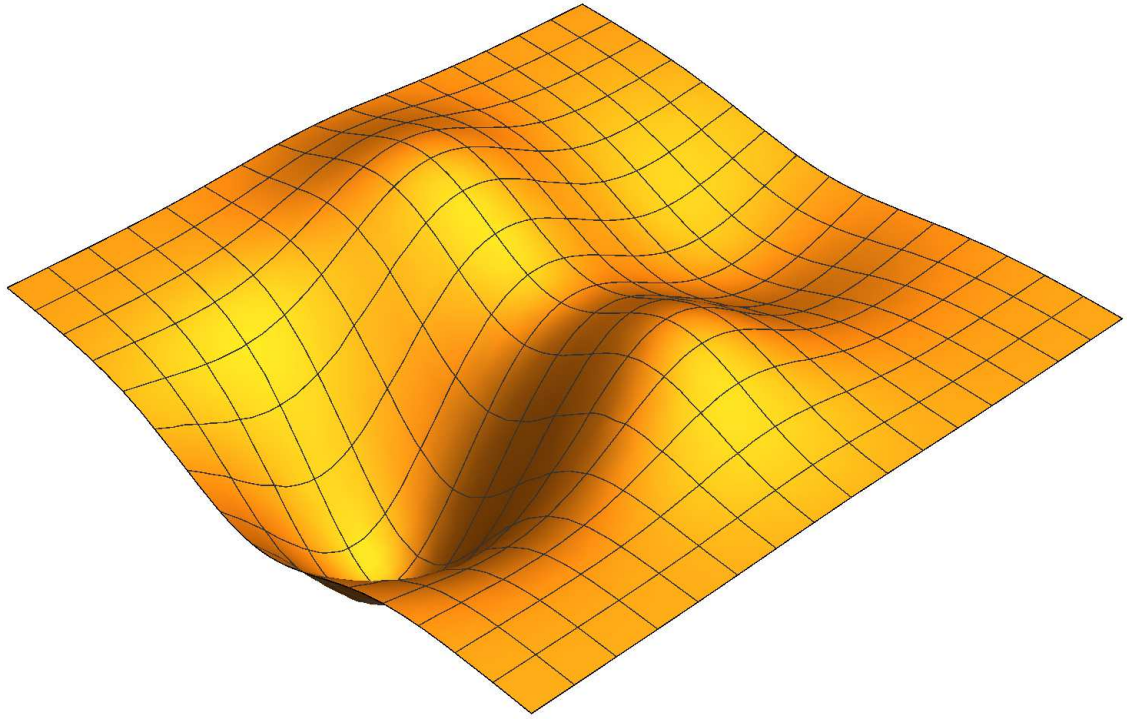
Convex: converges to global minimum; **dimension-free** iterations.



Nonconvex Optimization

Non-convex: converges to Stationary Point (SP) $\nabla f(\mathbf{x}) = 0$.

SP : local min / local max / saddle points



Many applications: no spurious local min (see full list later).

Some Well-Behaved Nonconvex Problems

- PCA, CCA, Matrix Factorization
- Orthogonal Tensor Decomposition (Ge, Huang, Jin, Yang, 2015)
- Complete Dictionary Learning (Sun et al, 2015)
- Phase Retrieval (Sun et al, 2015)
- Matrix Sensing (Bhojanapalli et al, 2016; Park et al, 2016)
- Symmetric Matrix Completion (Ge et al, 2016)
- Matrix Sensing/Completion, Robust PCA (Ge, Jin, Zheng, 2017)

- The problems have no spurious local minima and all saddle points are strict

Convergence to FOSP

Function $f(\cdot)$ is ℓ-**smooth (or gradient Lipschitz)**

$$\forall \mathbf{x}_1, \mathbf{x}_2, \ \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq \ell \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Point $\mathbf{x}$ is an ϵ-**first-order stationary point** (ϵ-FOSP) if

$$\|\nabla f(\mathbf{x})\| \leq \epsilon$$

Convergence to FOSP

Function $f(\cdot)$ is ℓ-**smooth (or gradient Lipschitz)**

$$\forall \mathbf{x}_1, \mathbf{x}_2, \ \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq \ell \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Point $\mathbf{x}$ is an ϵ-**first-order stationary point** (ϵ-FOSP) if

$$\|\nabla f(\mathbf{x})\| \leq \epsilon$$

Theorem [GD Converges to FOSP (Nesterov, 1998)]

For ℓ-smooth function, GD with $\eta = 1/\ell$ finds ϵ-FOSP in iterations:

$$\frac{2\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2}$$

*Number of iterations is dimension free.

Definitions and Algorithm

Function $f(\cdot)$ is ρ-**Hessian Lipschitz** if

$$\forall \mathbf{x}_1, \mathbf{x}_2, \ \|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\| \leq \rho \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Point $\mathbf{x}$ is an ϵ-**second-order stationary point** (ϵ-SOSP) if

$$\|\nabla f(\mathbf{x})\| \leq \epsilon, \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\sqrt{\rho\epsilon}$$

Definitions and Algorithm

Function $f(\cdot)$ is ρ-**Hessian Lipschitz** if

$$\forall \mathbf{x}_1, \mathbf{x}_2, \ \|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\| \leq \rho \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Point $\mathbf{x}$ is an ϵ-**second-order stationary point** (ϵ-SOSP) if

$$\|\nabla f(\mathbf{x})\| \leq \epsilon, \qquad \text{and} \qquad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\sqrt{\rho\epsilon}$$

Algorithm Perturbed Gradient Descent (PGD)

1. **for** $t = 0, 1, \ldots$ **do**
2. **if** perturbation condition holds **then**
3. $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t, \qquad \xi_t$ uniformly $\sim \mathbb{B}_0(r)$
4. $\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \nabla f(\mathbf{x}_t)$

Adds perturbation when $\|\nabla f(\mathbf{x}_t)\| \leq \epsilon$; no more than once per T steps.

Main Result

Theorem [PGD Converges to SOSP]

For ℓ-smooth and ρ-Hessian Lipschitz function f, PGD with $\eta = O(1/\ell)$ and proper choice of r, T w.h.p. finds ϵ-SOSP in iterations:

$$\tilde{O}\left(\frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2}\right)$$

*Dimension dependence in iteration is $\log^4(d)$ (almost dimension free).

Main Result

Theorem [PGD Converges to SOSP]

For ℓ-smooth and ρ-Hessian Lipschitz function f, PGD with $\eta = O(1/\ell)$ and proper choice of r, T w.h.p. finds ϵ-SOSP in iterations:

$$\tilde{O}\left(\frac{\ell(f(\mathbf{x}_0) - f^\star)}{\epsilon^2}\right)$$

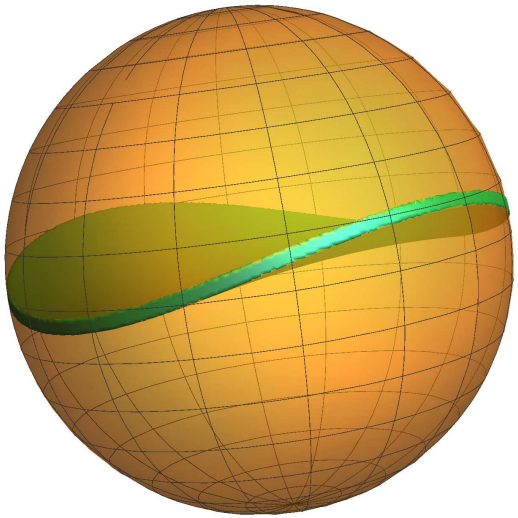
*Dimension dependence in iteration is $\log^4(d)$ (almost dimension free).

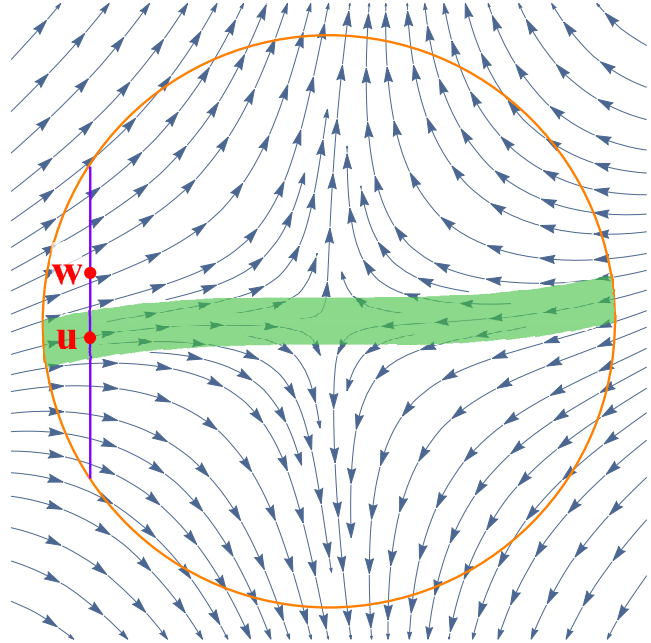
	GD(Nesterov 1998)	**PGD**(This Work)
Assumptions	ℓ-grad-Lip	ℓ-grad-Lip + ρ-Hessian-Lip
Guarantees	ϵ-FOSP	ϵ-SOSP
Iterations	$2\ell(f(\mathbf{x}_0) - f^\star)/\epsilon^2$	$\tilde{O}(\ell(f(\mathbf{x}_0) - f^\star)/\epsilon^2)$

Geometry and Dynamics around Saddle Points

Challenge: non-constant Hessian + large step size $\eta = O(1/\ell)$.

Around saddle point, **stuck region** forms a non-flat "pancake" shape.

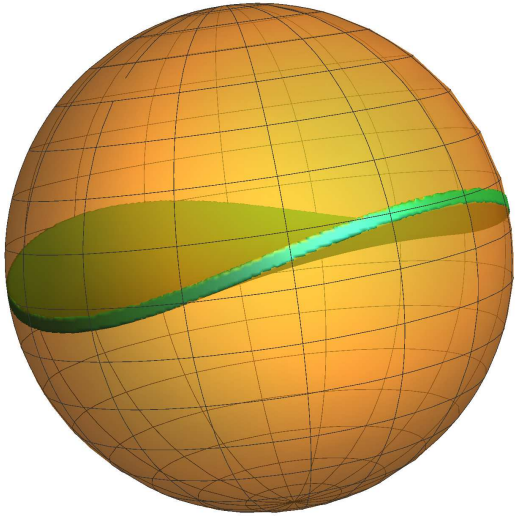


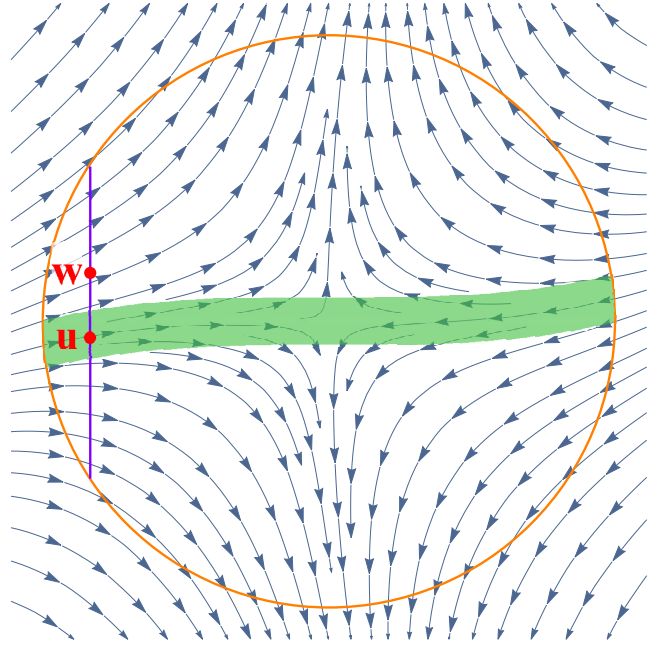


Geometry and Dynamics around Saddle Points

Challenge: non-constant Hessian + large step size $\eta = O(1/\ell)$.

Around saddle point, **stuck region** forms a non-flat "pancake" shape.





Key Observation: although we don't know its shape, we know it's thin!
(Based on an analysis of two nearly coupled sequences)

Next Questions

- Does acceleration help in escaping saddle points?
- What other kind of stochastic models can we use to escape saddle points?
- How do acceleration and stochastics interact?

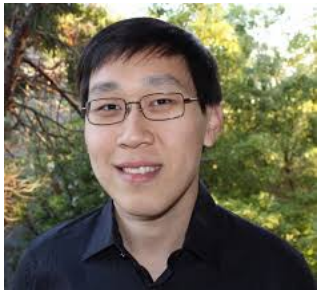
Next Questions

- Does acceleration help in escaping saddle points?
- What other kind of stochastic models can we use to escape saddle points?
- How do acceleration and stochastics interact?

- To address these questions we need to understand develop a deeper understanding of acceleration than has been available in the literature to date

Part II: Variational, Hamiltonian and Symplectic Perspectives on Acceleration

with Andre Wibisono, Ashia Wilson and Michael Betancourt



Interplay between Differentiation and Integration

- The 300-yr-old fields: Physics, Statistics
  - cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions
- The numerical disciplines
  - e.g.,. finite elements, Monte Carlo

Interplay between Differentiation and Integration

- The 300-yr-old fields: Physics, Statistics
  - cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions
- The numerical disciplines
  - e.g.,. finite elements, Monte Carlo
- Optimization?

Interplay between Differentiation and Integration

- The 300-yr-old fields: Physics, Statistics
  - cf. Lagrange/Hamilton, Laplace expansions, saddlepoint expansions
- The numerical disciplines
  - e.g.,. finite elements, Monte Carlo
- Optimization?
  - to date, almost entirely focused on differentiation

Accelerated gradient descent

Setting: Unconstrained convex optimization

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

- Classical gradient descent:

$$x_{k+1} = x_k - \beta \nabla f(x_k)$$

obtains a convergence rate of $O(1/k)$

Accelerated gradient descent

Setting: Unconstrained convex optimization

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

- Classical gradient descent:

$$x_{k+1} = x_k - \beta \nabla f(x_k)$$

  obtains a convergence rate of $O(1/k)$

- Accelerated gradient descent:

$$\begin{aligned} y_{k+1} &= x_k - \beta \nabla f(x_k) \\ x_{k+1} &= (1 - \lambda_k) y_{k+1} + \lambda_k y_k \end{aligned}$$

  obtains the (optimal) convergence rate of $O(1/k^2)$

The acceleration phenomenon

Two classes of algorithms:

- **Gradient methods**
  - Gradient descent, mirror descent, cubic-regularized Newton's method (Nesterov and Polyak '06), etc.
  - Greedy descent methods, relatively well-understood

The acceleration phenomenon

Two classes of algorithms:

- **Gradient methods**
  - Gradient descent, mirror descent, cubic-regularized Newton's method (Nesterov and Polyak '06), etc.
  - Greedy descent methods, relatively well-understood
- **Accelerated methods**
  - Nesterov's accelerated gradient descent, accelerated mirror descent, accelerated cubic-regularized Newton's method (Nesterov '08), etc.
  - Important for both theory (optimal rate for first-order methods) and practice (many extensions: FISTA, stochastic setting, etc.)
  - *Not* descent methods, faster than gradient methods, still mysterious

Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow

$$\dot{X}_t = -\nabla f(X_t)$$

(and mirror descent is discretization of natural gradient flow)

Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow

$$\dot{X}_t = -\nabla f(X_t)$$

(and mirror descent is discretization of natural gradient flow)

- Su, Boyd, Candes '14: Continuous time limit of accelerated gradient descent is a second-order ODE

$$\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0$$

Accelerated methods: Continuous time perspective

- Gradient descent is discretization of gradient flow

$$\dot{X}_t = -\nabla f(X_t)$$

(and mirror descent is discretization of natural gradient flow)

- Su, Boyd, Candes '14: Continuous time limit of accelerated gradient descent is a second-order ODE

$$\ddot{X}_t + \frac{3}{t}\dot{X}_t + \nabla f(X_t) = 0$$

- These ODEs are obtained by taking continuous time limits. Is there a deeper generative mechanism?

Our work: A general variational approach to acceleration

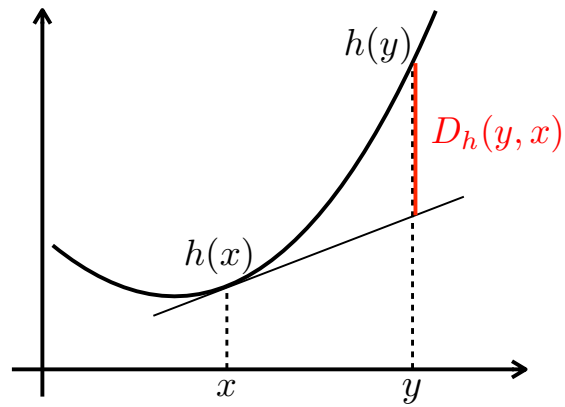
A systematic discretization methodology

Bregman Lagrangian

Define the **Bregman Lagrangian**:

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(D_h(x + e^{-\alpha_t} \dot{x}, x) - e^{\beta_t} f(x) \right)$$

- Function of position x, velocity $\dot{x}$, and time t
- $D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle$ is the Bregman divergence
- h is the convex distance-generating function
- f is the convex objective function

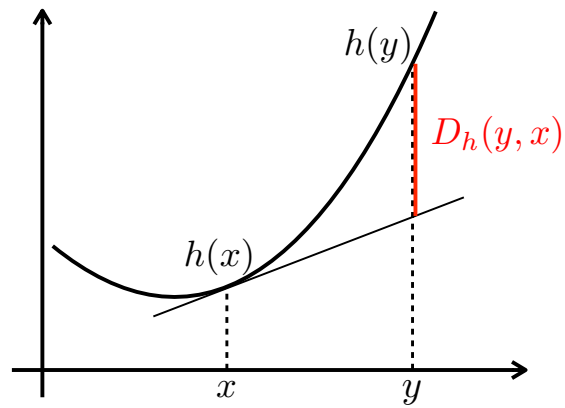


Bregman Lagrangian

Define the **Bregman Lagrangian**:

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t - \alpha_t} \left(\frac{1}{2} \|\dot{x}\|^2 - e^{2\alpha_t + \beta_t} f(x) \right)$$

- Function of position x, velocity $\dot{x}$, and time t
- $D_h(y, x) = h(y) - h(x) - \langle \nabla h(x), y - x \rangle$ is the Bregman divergence
- h is the convex distance-generating function
- f is the convex objective function
- $\alpha_t, \beta_t, \gamma_t \in \mathbb{R}$ are arbitrary smooth functions
- In Euclidean setting, simplifies to damped Lagrangian

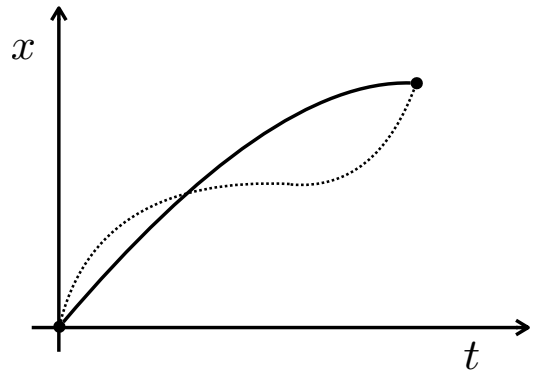


Bregman Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(D_h(x + e^{-\alpha_t} \dot{x}, x) - e^{\beta_t} f(x) \right)$$

Variational problem over curves:

$$\min_X \int \mathcal{L}(X_t, \dot{X}_t, t)\, dt$$



Optimal curve is characterized by **Euler-Lagrange** equation:

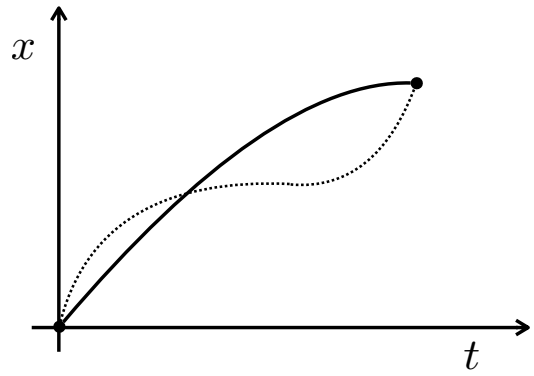
$$\frac{d}{dt} \left\{ \frac{\partial \mathcal{L}}{\partial \dot{x}}(X_t, \dot{X}_t, t) \right\} = \frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t)$$

Bregman Lagrangian

$$\mathcal{L}(x, \dot{x}, t) = e^{\gamma_t + \alpha_t} \left(D_h(x + e^{-\alpha_t} \dot{x}, x) - e^{\beta_t} f(x) \right)$$

Variational problem over curves:

$$\min_X \int \mathcal{L}(X_t, \dot{X}_t, t)\, dt$$



Optimal curve is characterized by **Euler-Lagrange** equation:

$$\frac{d}{dt} \left\{ \frac{\partial \mathcal{L}}{\partial \dot{x}}(X_t, \dot{X}_t, t) \right\} = \frac{\partial \mathcal{L}}{\partial x}(X_t, \dot{X}_t, t)$$

E-L equation for Bregman Lagrangian under ideal scaling:

$$\ddot{X}_t + (e^{\alpha_t} - \dot{\alpha}_t)\dot{X}_t + e^{2\alpha_t + \beta_t} \left[\nabla^2 h(X_t + e^{-\alpha_t} \dot{X}_t) \right]^{-1} \nabla f(X_t) = 0$$

General convergence rate

Theorem

Theorem Under ideal scaling, the E-L equation has convergence rate

$$f(X_t) - f(x^*) \leq O(e^{-\beta_t})$$

Proof. Exhibit a Lyapunov function for the dynamics:

$$\mathcal{E}_t = D_h\left(x^*,\, X_t + e^{-\alpha_t}\dot{X}_t\right) + e^{\beta_t}(f(X_t) - f(x^*))$$

$$\dot{\mathcal{E}}_t = -e^{\alpha_t+\beta_t} D_f(x^*, X_t) + (\dot{\beta}_t - e^{\alpha_t})e^{\beta_t}(f(X_t) - f(x^*)) \;\leq\; 0$$

□

Note: Only requires convexity and differentiability of f, h

Mysteries

- Why can't we discretize the dynamics when we are using exponentially fast clocks?
- What happens when we arrive at a clock speed that we can discretize?
- How do we discretize once it's possible?

Symplectic Integration

- Consider discretizing a system of differential equations obtained from physical principles
- Solutions of the differential equations generally conserve various quantities (energy, momentum, volumes in phase space)
- Is it possible to find discretizations whose solutions exactly conserve these same quantities?
- Yes!
  - from a long line of research initiated by Jacobi, Hamilton, Poincare' and others

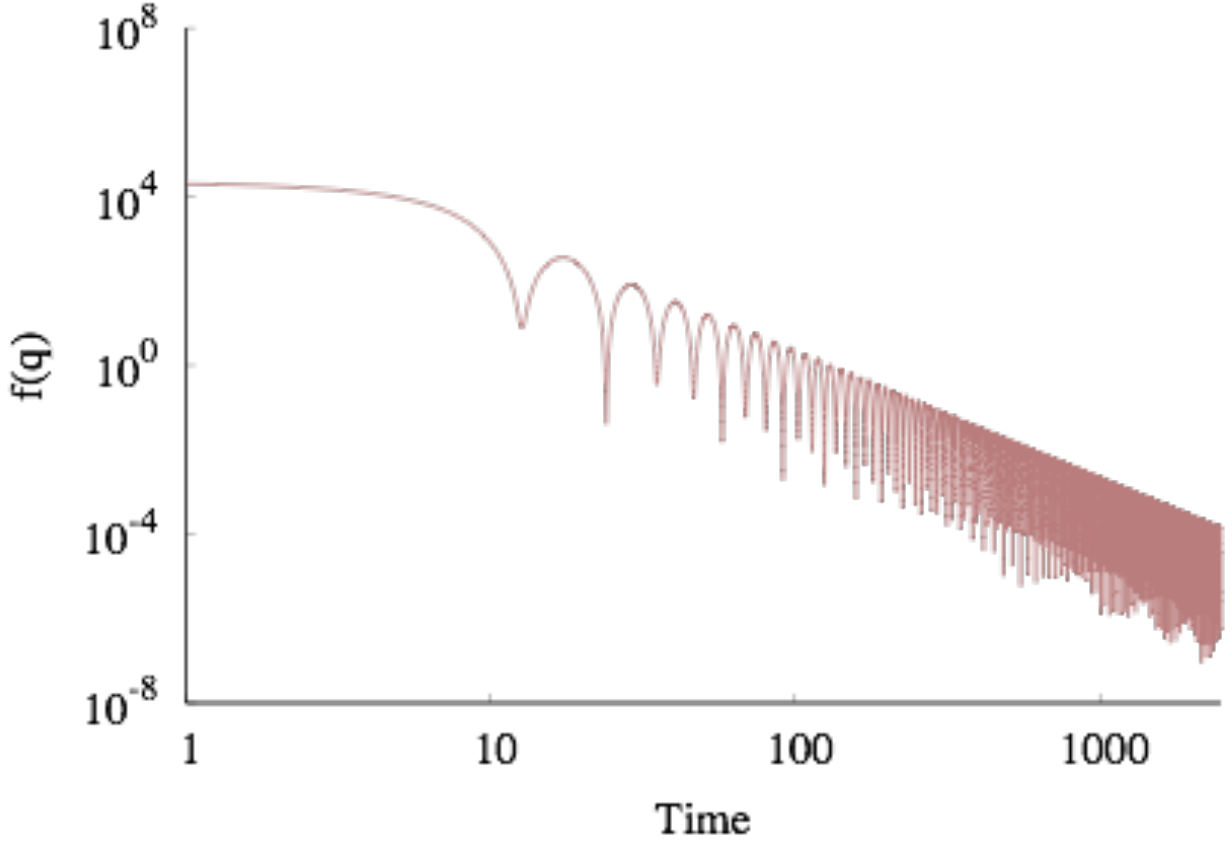
Towards A Symplectic Perspective

- We've discussed discretization of Lagrangian-based dynamics
- Discretization of Lagrangian dynamics is often fragile and requires small step sizes
- We can build more robust solutions by taking a Legendre transform and considering a *Hamiltonian* formalism:

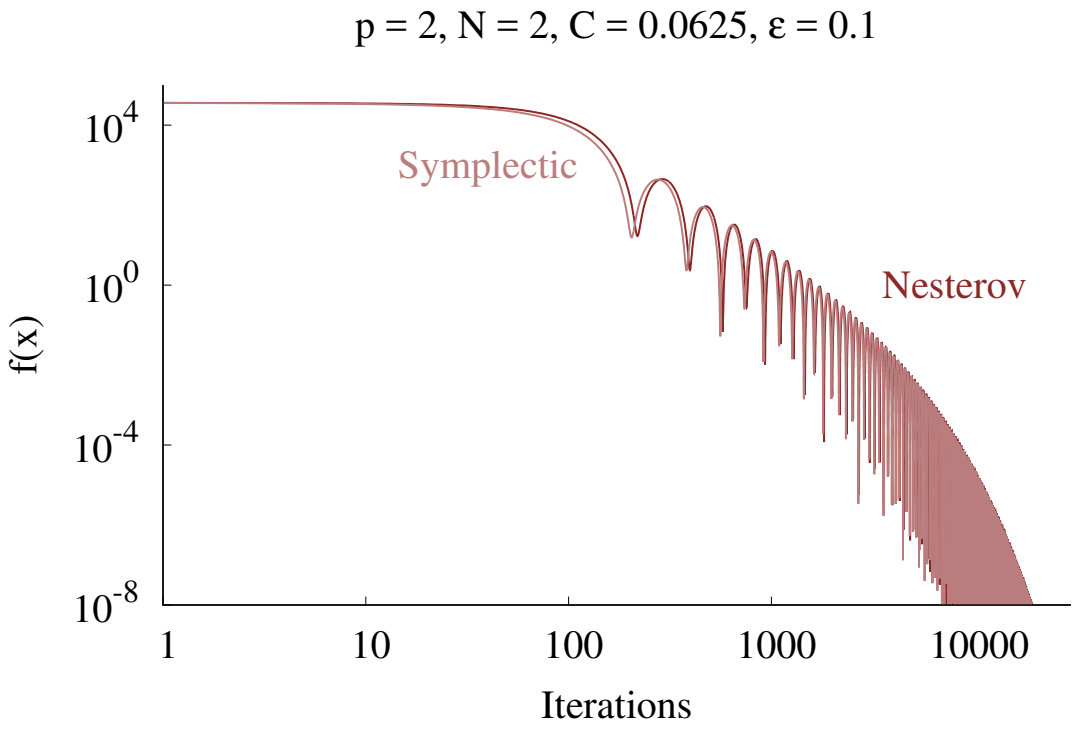
$$L(q, v, t) \rightarrow H(q, p, t, \mathcal{E})$$

$$\left(\frac{\mathrm{d}q}{\mathrm{d}t}, \frac{\mathrm{d}v}{\mathrm{d}t}\right) \rightarrow \left(\frac{\mathrm{d}q}{\mathrm{d}\tau}, \frac{\mathrm{d}p}{\mathrm{d}\tau}, \frac{\mathrm{d}t}{\mathrm{d}\tau}, \frac{\mathrm{d}\mathcal{E}}{\mathrm{d}\tau}\right)$$

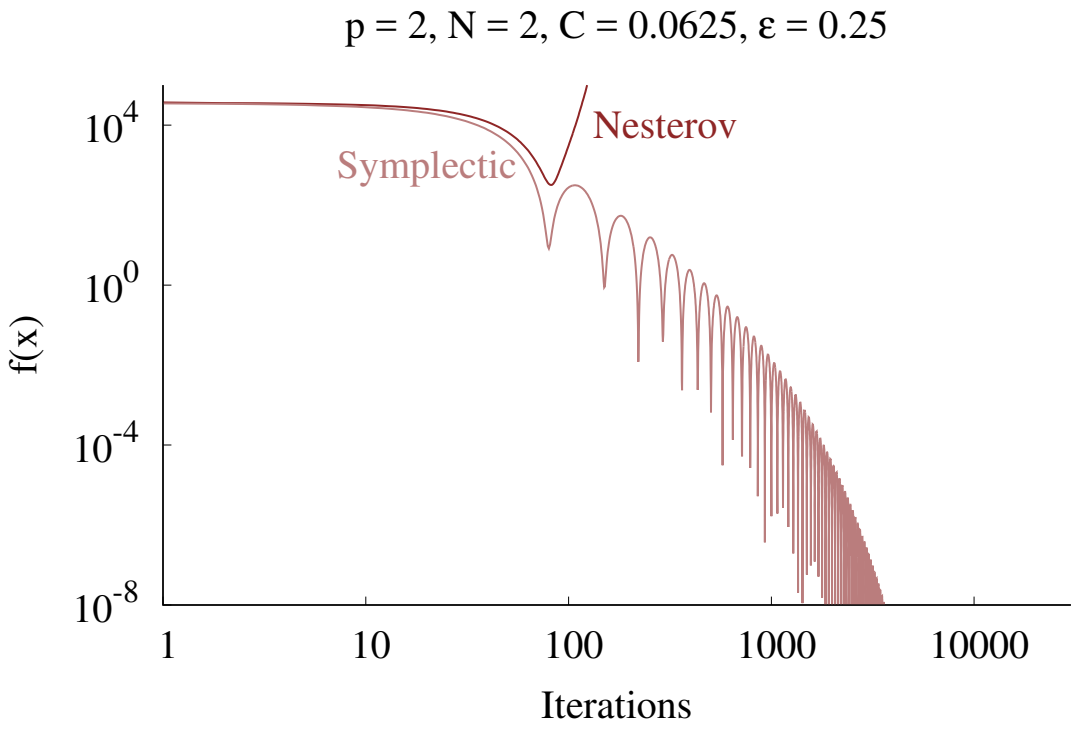
Symplectic Integration of Bregman Hamiltonian



Symplectic vs Nesterov



Symplectic vs Nesterov



Part III: Acceleration and Saddle Points

with Chi Jin and Praneeth Netrapalli

Problem Setup

Smooth Assumption: $f(\cdot)$ is smooth:

- ℓ-gradient Lipschitz, i.e.

$$\forall \mathbf{x}_1, \mathbf{x}_2, \ \|\nabla f(\mathbf{x}_1) - \nabla f(\mathbf{x}_2)\| \leq \ell \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

- ρ-Hessian Lipschitz, i.e.

$$\forall \mathbf{x}_1, \mathbf{x}_2, \ \|\nabla^2 f(\mathbf{x}_1) - \nabla^2 f(\mathbf{x}_2)\| \leq \rho \|\mathbf{x}_1 - \mathbf{x}_2\|.$$

Goal: find **second-order stationary point** (SOSP):

$$\nabla f(\mathbf{x}) = 0, \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq 0.$$

Relaxed version: ϵ-**second-order stationary point** (ϵ-SOSP):

$$\|\nabla f(\mathbf{x})\| \leq \epsilon, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x})) \geq -\sqrt{\rho\epsilon}$$

Analysis of AGD in the Nonconvex Setting

- **Challenge**: AGD is not a descent algorithm
- **Solution**: Lift the problem to a phase space, and make use of a Hamiltonian
- **Consequence**: AGD is nearly a descent algorithm in the Hamiltonian, with a simple "negative curvature exploitation" (NCE; cf. Carmon et al., 2017) step handling the case when descent isn't guaranteed

Hamiltonian Perspective on AGD

- AGD is a discretization of the following ODE

$$\ddot{x} + \tilde{\theta}\dot{x} + \nabla f(x) = 0$$

- Multiplying by $\dot{x}$ and integrating from t_1 to t_2 gives us

$$f(x_{t_2}) + \frac{1}{2}\|\dot{x}_{t_2}\|^2 = f(x_{t_1}) + \frac{1}{2}\|\dot{x}_{t_1}\|^2 - \tilde{\theta}\int_{t_1}^{t_2}\|\dot{x}_t\|^2 dt$$

- In convex case, Hamiltonian $f(x_t) + \frac{1}{2}\|\dot{x}_t\|^2$ decreases monotonically

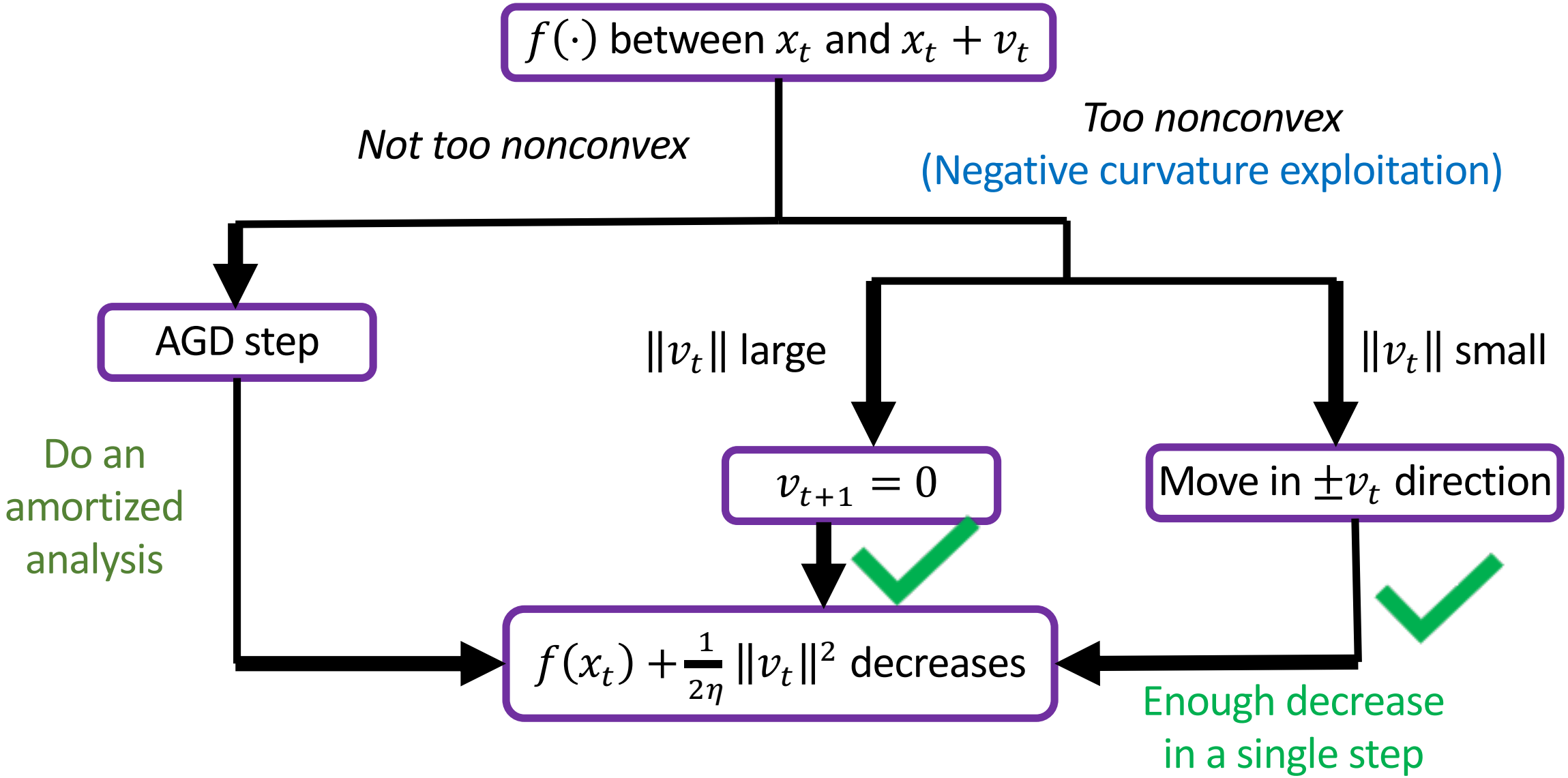
Algorithm

Algorithm Perturbed Accelerated Gradient Descent (PAGD)

1. **for** $t = 0, 1, \ldots$ **do**
2. **if** $\|\nabla f(\mathbf{x}_t)\| \leq \epsilon$ *and* no perturbation in last T steps **then**
3. $\mathbf{x}_t \leftarrow \mathbf{x}_t + \xi_t, \qquad \xi_t$ uniformly $\sim \mathbb{B}_0(r)$
4. $\mathbf{y}_t \leftarrow \mathbf{x}_t + (1-\theta)\mathbf{v}_t$
5. $\mathbf{x}_{t+1} \leftarrow \mathbf{y}_t - \eta \nabla f(\mathbf{y}_t); \quad \mathbf{v}_{t+1} \leftarrow \mathbf{x}_{t+1} - \mathbf{x}_t$
6. **if** $f(\mathbf{x}_t) \leq f(\mathbf{y}_t) + \langle \nabla f(\mathbf{y}_t), \mathbf{x}_t - \mathbf{y}_t \rangle - \frac{\gamma}{2}\|\mathbf{x}_t - \mathbf{y}_t\|^2$ **then**
7. $\mathbf{x}_{t+1} \leftarrow \mathrm{NCE}(\mathbf{x}_t, \mathbf{v}_t, s); \quad \mathbf{v}_{t+1} \leftarrow 0$

- Perturbation (line 2-3);
- Standard AGD (line 4-5);
- Negative Curvature Exploitation (NCE, line 6-7)
  - 1) simple (two steps), 2) auxiliary. [inspired by Carmon et al. 2017]

Hamiltonian Analysis



Convergence Result

PAGD Converges to SOSP Faster (Jin et al. 2017)

For ℓ-gradient Lipschitz and ρ-Hessian Lipschitz function f, PAGD with proper choice of $\eta, \theta, r, T, \gamma, s$ w.h.p. finds ϵ-SOSP in iterations:

$$\tilde{O}\left(\frac{\ell^{1/2}\rho^{1/4}(f(\mathbf{x}_0)-f^\star)}{\epsilon^{7/4}}\right)$$

	Strongly Convex	**Nonconvex (SOSP)**
Assumptions	ℓ-grad-Lip & α-str-convex	ℓ-grad-Lip & ρ-Hessian-Lip
(Perturbed) GD	$\tilde{O}(\ell/\alpha)$	$\tilde{O}(\Delta_f \cdot \ell/\epsilon^2)$
(Perturbed) AGD	$\tilde{O}(\sqrt{\ell/\alpha})$	$\tilde{O}(\Delta_f \cdot \ell^{\frac{1}{2}}\rho^{\frac{1}{4}}/\epsilon^{\frac{7}{4}})$
Condition κ	ℓ/α	$\ell/\sqrt{\rho\epsilon}$
Improvement	$\sqrt{\kappa}$	$\sqrt{\kappa}$

Part IV: Acceleration and Stochastics

with Xiang Cheng, Niladri Chatterji and Peter Bartlett

Acceleration and Stochastics

- Can we accelerate diffusions?
- There have been negative results...

Acceleration and Stochastics

- Can we accelerate diffusions?
- There have been negative results…
- …but they've focused on classical overdamped diffusions

Acceleration and Stochastics

- Can we accelerate diffusions?
- There have been negative results…
- …but they've focused on classical overdamped diffusions
- Inspired by our work on acceleration, can we accelerate underdamped diffusions?

Overdamped Langevin MCMC

Described by the Stochastic Differential Equation (SDE):

$$dx_t = -\nabla U(x_t)dt + \sqrt{2}dB_t$$

where $U(x)$: $R^d \to R$ and B_t is standard Brownian motion.

The stationary distribution is $p^*(x) \propto \exp\big(U(x)\big)$

Corresponding Markov Chain Monte Carlo Algorithm (MCMC):

$$\tilde{x}_{(k+1)\delta} = \tilde{x}_{k\delta} - \nabla U(\tilde{x}_{k\delta}) + \sqrt{2\delta}\xi_k$$

where δ is the *step-size* and $\xi_k \sim N(0, I_{d\times d})$

Guarantees under Convexity

Assuming $U(x)$ is L-smooth and m-strongly convex:

Dalalyan'14: Guarantees in Total Variation

If $n \geq O\left(\frac{d}{\epsilon^2}\right)$ then, $TV(p^{(n)}, p^*) \leq \epsilon$

Durmus & Moulines'16: Guarantees in 2-Wasserstein

If $n \geq O\left(\frac{d}{\epsilon^2}\right)$ then, $W_2(p^{(n)}, p^*) \leq \epsilon$

Cheng and Bartlett'17: Guarantees in KL divergence

If $n \geq O\left(\frac{d}{\epsilon^2}\right)$ then, $\mathrm{KL}(p^{(n)}, p^*) \leq \epsilon$

Underdamped Langevin Diffusion

Described by the *second-order* equation:

$$dx_t = v_t dt$$
$$dv_t = -\gamma v_t dt + \lambda \nabla U(x_t) dt + \sqrt{2\gamma\lambda}\, dB_t$$

The stationary distribution is $p^*(x, v) \propto \exp\left(-U(x) - \frac{|v|_2^2}{2\lambda}\right)$

Intuitively, x_t is the position and v_t is the velocity

$\nabla U(x_t)$ is the force and γ is the drag coefficient

Discretization

We can discretize; and at each step evolve according to

$$d\tilde{x}_t = \tilde{v}_t dt$$
$$d\tilde{v}_t = -\gamma \tilde{v}_t dt \; - \lambda \nabla U\left(\tilde{x}_{\lfloor t/\delta \rfloor \delta}\right) dt + \sqrt{2\gamma\lambda}\, dB_t$$

we evolve this for time δ to get an MCMC algorithm

Notice this is a *second-order* method. Can we get faster rates?

Quadratic Improvement

Let $p^{(n)}$ denote the distribution of $(\tilde{x}_{n\delta}, \tilde{v}_{n\delta})$. Assume $U(x)$ is strongly convex

Cheng, Chatterji, Bartlett, Jordan '17:

If $n \geq O\left(\frac{\sqrt{d}}{\epsilon}\right)$ then $W_2\left(p^{(n)}, p^*\right) \leq \epsilon$

Compare with Durmus & Moulines '16 (Overdamped)

If $n \geq O\left(\frac{d}{\epsilon^2}\right)$ then $W_2\left(p^{(n)}, p^*\right) \leq \epsilon$

On Dissipative Symplectic Integration with Applications to Gradient-Based Optimization

Guiherme França, Michael I. Jordan, and René Vidal

based on arXiv:2004.06840 [math.OC]

Motivation

Suppose we have a **dissipative** Hamiltonian system:

$$\frac{dq^j}{dt} = \frac{\partial H}{\partial p_j}, \qquad \frac{dp_j}{dt} = -\frac{\partial H}{\partial q^j}, \qquad H = H(t, q, p),$$

where $q \in \mathcal{M}^n$ (smooth manifold) and $(q, p) \in T^*\mathcal{M}$ (cotangent bundle) $(j = 1, \ldots, n)$. Assume that its trajectories can be viewed as solving

$$\min_{q \in \mathcal{M}} f(q),$$

and that we understand the dynamics; i.e. stability, convergence rates, etc. A fundamental question is the following:

- **Which discretizations are able to preserve the stability and rates of convergence of such a continuous-time system?**
- The answer would give us a systematic way to derive efficient optimization algorithms ("acceleration") ...
- ... without the need for a discrete-time convergence analysis.

Motivation

Conservative Hamiltonian systems are ubiquitous — $H(q, p)$ is independent of time. But the conservation of energy precludes convergence to a point; consider the harmonic oscillator.

This is not what we want in optimization.
We need "dissipation" — where $H(t, q, p)$ is explicitly time-dependent — which leads us to another important question:

- **Can we map optimization algorithms into dissipative continuous-time dynamical systems that provide analytical insight into the behavior of the algorithm?**
- The answer would allow us to infer stability and convergence rates of such algorithms with a broader mathematical machinery than traditionally available.

Our approach

- Those two questions are related. The ability to **preserve** convergence rates can be seen as some kind of **"invariance."**
- But a dissipative system presumably has no conservation law.
- Using **symplectic geometry**, we will show that a dissipative Hamiltonian system can be seen as a conservative Hamiltonian system in higher dimensions (symplectification + gauge fixing).
- Together with **backward-error analysis** we can bring these ideas to discrete-time to obtain a framework (presymplectic integrators) where the stability and convergence rates of the continuous system are preserved (up to a small and controlled error).

Backward-error analysis

Consider a dynamical system over a smooth manifold $\mathcal{M}$:

$$\dot{x}(t) = X(x(t)),$$

where X is the vector field and $\varphi_t = e^{tX}$ is its **flow map**.

A **numerical map** ϕ_h, of **order** $r \geq 1$, is an approximation ($h > 0$):

$$\|\phi_h(x) - \varphi_h(x)\| = \mathcal{O}(h^{r+1}) \qquad \text{for any } x \in \mathcal{M}.$$

Theorem

Every numerical method, ϕ_h, can be seen as the "exact flow" of a perturbed dynamical system:

$$\dot{x}(t) = \tilde{X}(x(t)), \qquad \tilde{X} = X + \Delta X_1 h + \Delta X_2 h^2 + \cdots .$$

These ideas have been developed since the late 90's in numerical analysis (Benettin, Giorgilli, Hairer, Reich, Lubich, ...).

Backward-error analysis

The perturbed vector field $\tilde{X}$ has to be **truncated**. Denoting by $\varphi_{t,\tilde{X}} = e^{t\tilde{X}}$ the associated flow, one has:

Theorem (Benettin, Giorgilli, Hairer, Reich, ...)

There exists a family of (truncated) perturbed vector fields, $\|X(x) - \tilde{X}(x)\| = \mathcal{O}(h^r)$, *such that* $\|\phi_h(x) - \varphi_{h,\tilde{X}}(x)\| \leq Che^{-r}e^{-h_0/h}$.

- This tells us that the numerical flow is very close to the "perturbed flow" (exponentially small error).
- For typical numerical integrators this result is not very useful. One is rather interested in comparing ϕ_h to φ_h (not $\varphi_{h,\tilde{X}}$).
- However, this result becomes **extremely useful** if one can show that $\tilde{X}$ has the "same structure" as X. This is why **structure-preserving** methods are special; e.g., symplectic integrators.

Symplectic manifolds and conservative Hamiltonians

Definition

An even-dimensional smooth manifold $\mathcal{M}$ endowed with a closed nondegenerate 2-form ω is a **symplectic manifold**.[a]

[a] ω maps two vectors into a number and it is a totally *skew-symmetric* object, $\omega(X, Y) = -\omega(Y, X)$, thus it imposes a special geometry on $\mathcal{M}$.

As an analogy, in going from the real to complex numbers one introduces $i^2 = -1$. Here, in a matrix representation, one introduces $\omega^2 = -I$ over $\mathcal{M}$.

- Symplectic geometry arises in several areas: classical mechanics, complex geometry, Lie groups and algebras, representation varieties, geometric quantization, and so on. They are worth studying in their own right and have a beautiful mathematical structure.

Symplectic manifolds and conservative Hamiltonians

The universality of symplectic manifolds and Hamiltonian systems follow from the following facts.

Theorem

The tangent bundle[a] $T^*\mathcal{M}$ *of any differentiable manifold* $\mathcal{M}$*, with coordinates* $q^1, \ldots, q^n, p_1, \ldots, p_n$*, is a symplectic manifold. The symplectic 2-form is given by* $\omega = \sum_j dp_j \wedge dq^j$*.*

[a] The tangent bundle is just the collection of all cotangent spaces, i.e. the collection of all (tensor products of) dual vector spaces.

Theorem

A dynamical system with phase space $T^*\mathcal{M}$ *preserves the simplectic structure* ω *if and only if it is a conservative Hamiltonian system.*[a]

[a] One with a time-independent Hamiltonian $H = H(q, p)$.

Symplectic manifolds and conservative Hamiltonians

1. By "preserving" we mean that the Lie derivative of the Hamiltonian vector field obeys $\mathcal{L}_{X_H}\omega = 0$. This is the first fundamental property.
2. The second fundamental property is energy conservation: $\frac{dH}{dt} = 0$.

Definition

It is possible to construct a class of numerical integrators, ϕ_h, that exactly preserve ω: $\phi_h^* \circ \omega \circ \phi_h = \omega$. They are called **symplectic integrators**.

1. This implies that the perturbed dynamical system associated to ϕ_h obeys $\mathcal{L}_{\tilde{X}}\omega = 0$.[1] Thus, X_H and $\tilde{X}$ have the same structure!
2. The last theorem above implies that the perturbed system must be a Hamiltonian system, with a perturbed $\tilde{H}$, and for which $\frac{d\tilde{H}}{dt} = 0$.

[1] Recall that $\tilde{X}$ is the vector field of the perturbed system, associated to ϕ_h.

Why symplectic integrators are so successful?

We can now use the previous general backward-error analysis theorem:

Theorem (Benettin, Giorgilli)

Let ϕ_h be a symplectic integrator of order r. Assume H is Lipschitz. Then for large simulation times $t_\ell = h\ell = \mathcal{O}(h^r e^r e^{h_0/h})$, $\ell = 0, 1, \ldots,$ we have

$$\underbrace{H \circ \phi_{t_\ell}}_{\text{discrete}} = \underbrace{H \circ \varphi_{t_\ell}}_{\text{continuous}} + \underbrace{\mathcal{O}(h^r)}_{\text{bounded error}}$$

1. A symplectic integrator preserves the symplectic form, ω, exactly;
2. It "almost" preserves the energy, H (up to a bounded error).

however . . . things break down in a dissipative setting!

There is one crucial assumption behind all of this: *the Hamiltonian is a constant of motion $H = const.$* Therefore, these arguments break down when H varies over time, i.e., in the absence of a conservation law.

Dissipative Hamiltonian systems

- Since $H(t, q, p)$ depends on time, the Hamiltonian is not conserved:

$$\frac{dH}{dt} = \frac{\partial H}{\partial t} \neq 0.$$

- One can also show that the symplectic form is no longer preserved, $\mathcal{L}_{X_H}\omega \neq 0$. Thus, the phase space is no longer a symplectic manifold.
- One can "naively" apply a symplectic integrator to a dissipative system, but there is no existing result that extends that "main theorem"—close preservation of H and long term stability—into a dissipative setting ...
- ...What is the geometry of the phase space? Does the numerical method reproduces the Hamiltonian? Does it has long time stability?

Symplectification

There is a generalization of symplectic manifolds:

Definition

A **presymplectic manifold** $\mathcal{M}$ has dimension $2n+\bar{n}$ ($\bar{n} \geq 0$), and a 2-form ω of rank $2n$ everywhere. (The presymplectic form ω is degenerate.)[a]

[a] In our case, $\bar{n} = 1$.

It is possible to construct a **conservative** Hamiltonian system, $\mathscr{H}$, on a higher-dimensional **symplectic manifold** $T^*\hat{\mathcal{M}}$, of dimension $2n+2$. Let its coordinates be (q^μ, p_μ), for $\mu = 0, 1, \ldots, n$:

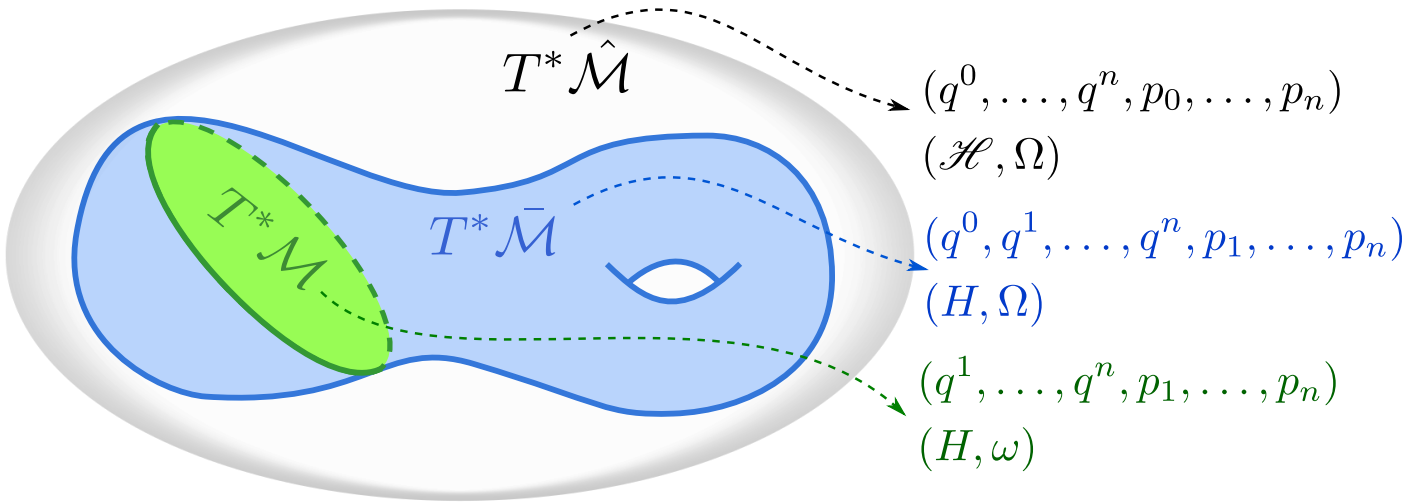
$$\frac{dq^\mu}{ds} = \frac{\partial \mathscr{H}}{\partial p_\mu}, \qquad \frac{dp_\mu}{ds} = -\frac{\partial \mathscr{H}}{\partial q^\mu}, \qquad \frac{d\mathscr{H}}{ds} = 0 \quad \text{(energy conservation)}.$$

Here s is a "new time parameter." Then it is possible to **embed** the original dissipative system into this symplectic manifold.

Symplectification

By removing the spurious degrees of freedom—gauge fixing—i.e., setting $q^0 = t = s$ and $p_0 = H(s) \equiv H(q(s), p(s))$—this is a function of time—the dissipative system lies on a hypersurface $\mathscr{H} = \text{const.}$ defined by:

$$\mathscr{H}(q^0, \ldots, q^n, p_0, \ldots, p_n) = p_0(s) + H(q^0, q^1, \ldots, q^n, p_1, \ldots, p_n).$$



Under this correspondence, the symplectic structure of the higher dimensional conservative system, Ω, recovers the "presymplectic structure" of the dissipative system, ω.

Presymplectic integrators

We define the following class of numerical methods:

Definition

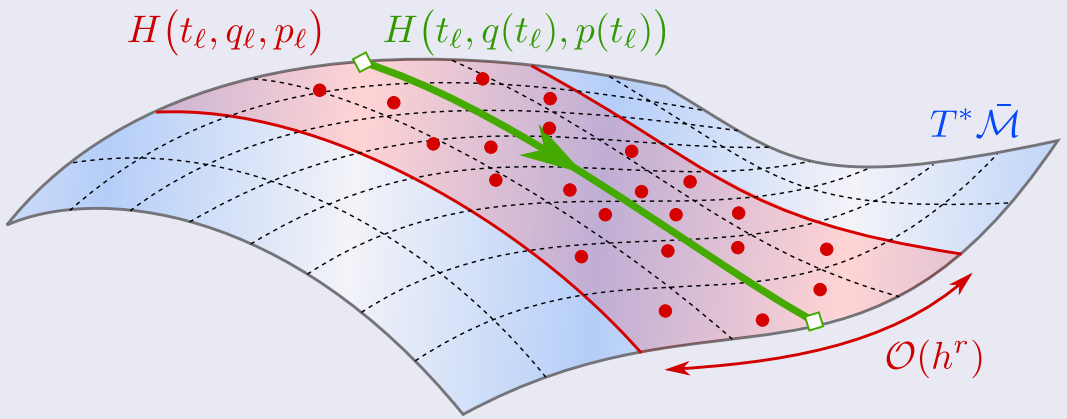
ϕ_h is a **presymplectic integrator** for the **dissipative** Hamiltonian system if it is a **reduction** of a symplectic integrator for its symplectification.

Theorem

Due to this correspondence, we can extend the range of standard theorems into a dissipative setting, where there is no conservation law. In particular, we can prove that the **decaying Hamiltonian is "preserved:"**

$$\underbrace{H \circ \phi_{t_\ell}}_{\text{numerical}} = \underbrace{H \circ \varphi_{t_\ell}}_{\text{continuous}} + \underbrace{\mathcal{O}(h^r)}_{\text{small error}}$$

$$\text{for } t_\ell \equiv h\ell = \mathcal{O}(h^r e^r e^{h_0/h})$$



Implications for optimization

We consider dissipative systems arising from the general class of Hamiltonians:

$$H \equiv e^{-\eta_1(t)} T(t, q, p) + e^{\eta_2(t)} f(q).$$

$\eta_1, \eta_2 \geq 0$ are increasing with t. In specific cases, we know how to obtain a continuous-time convergence rate: $f(q(t)) - f^\star \leq \mathcal{R}(t)$.

Corollary

A presymplectic integrator ϕ_h, of order $r \geq 1$, is a "rate-matching" discretization:

$$\underbrace{f(q_\ell) - f^\star}_{\text{discrete rate}} = \underbrace{f(q(t_\ell)) - f^\star}_{\text{continuous rate}} + \underbrace{\mathcal{O}\left(h^r e^{-\eta_2(t_\ell)}\right)}_{\text{tiny error}},$$

provided $e^{L_\phi t_\ell - \eta_1(t_\ell)} < \infty$ and for large $t_\ell \equiv h\ell = \mathcal{O}(h^r e^r e^{h_0/h})$.

Implications for optimization

- Under appropriate damping, presymplectic integrators can provide "rate-matching" discretizations.
- The error decreases with the order, $\sim h^r$, but is dominated by $\sim e^{-\eta_2(t)}$. Thus high-order integrators may not be necessary.
- If η_2 grows sufficiently fast, the error can be negligible; e.g. exponentially small.
- $\ell \sim h^{r-1} e^r e^{h_0/h}$ is astonishingly large; e.g., $h = 0.01$, $\ell \sim 10^{43}$.
- The strongest requirement is $e^{L_\phi t - \eta_1(t_\ell)} < \infty$, which "fixes" η_1. In particular, the "heavy ball damping", $\eta_1 = \gamma t$, or "Nesterov's damping", $\eta_1 = \gamma \log t$, can be seen as arising from this condition.
- Other choices may be possible, such as $\eta_1 = \gamma_1 \log t + \gamma_2 t^\delta$.

Example: the Bregman dynamics

The Bregman Hamiltonian provides a general approach to optimization (Wibisono, Wilson, MJ, PNAS 2016):

$$H = e^{\alpha+\gamma}\left\{D_{h^\star}\left(\nabla h(q) + e^{-\gamma}p, \nabla h(q)\right) + e^{\beta}f(q)\right\},$$

where D_h is the Bregman divergence, obtained in terms of a convex function $h(x)$, and $h^\star$ is its convex dual. Under appropriate "scaling conditions" on α, β, γ, Hamilton's equations are equivalent to

$$\ddot{q} + \left(e^{\alpha} - \dot{\alpha}\right)\dot{q} + e^{2\alpha+\beta}\left[\nabla^2 h\left(q + e^{-\alpha}\dot{q}\right)\right]^{-1}\nabla f(q) = 0.$$

For a convex function f, one can show that this system has a convergence rate given by:

$$f(q(t)) - f^\star = \mathcal{O}\left(e^{-\beta(t)}\right).$$

Bregman dynamics: separable case

Choosing $h(x) = \frac{1}{2} x \cdot Mx$, the kinetic energy simplifies and we have

$$H = \frac{1}{2} e^{-\eta_1(t)}\, p \cdot M^{-1} p + e^{\eta_2(t)} f(q), \quad \eta_1 \equiv \gamma - \alpha, \quad \eta_2 \equiv \alpha + \beta + \gamma.$$

One can now apply any presymplectic integrator (many possible choices are available). For instance, one based on the popular leapfrog method yields

$$\begin{aligned}
t_{\ell+1/2} &= t_\ell + h/2, \\
q_{\ell+1/2} &= q_\ell + (h/2) e^{-\eta_1(t_{\ell+1/2})} M^{-1} p_\ell, \\
p_{\ell+1} &= p_\ell - h e^{\eta_2(t_{\ell+1/2})} \nabla f(q_{\ell+1/2}), \\
t_{\ell+1} &= t_{\ell+1/2} + h/2, \\
q_{\ell+1} &= q_{\ell+1/2} + (h/2) e^{-\eta_1(t_{\ell+1/2})} M^{-1} p_{\ell+1}.
\end{aligned}$$

One can now make several choices for M, α, β, and γ to obtain a specific optimization algorithm that will respect the continuous convergence rate.

Bregman dynamics: nonseparable case

It is possible to construct **explicit** methods even though the general Bregman Hamiltonian is **nonseparable**. This is done by duplicating the degrees of freedom:

$$\bar{H}(t,q,p,\bar{t},\bar{q},\bar{p}) \equiv H(t,q,\bar{p}) + H(\bar{t},\bar{q},p) + \frac{\xi}{2}\left(\|q-\bar{q}\|^2 + \|p-\bar{p}\|^2\right).$$

We thus propose the following numerical maps:

$$\phi_h^A \begin{pmatrix} t \\ q \\ p \\ \bar{t} \\ \bar{q} \\ \bar{p} \end{pmatrix} = \begin{pmatrix} t \\ q \\ p - h\nabla_q H(t,q,\bar{p}) \\ \bar{t}+h \\ \bar{q} + h\nabla_{\bar{p}} H(t,q,\bar{p}) \\ \bar{p} \end{pmatrix}, \quad \phi_h^B \begin{pmatrix} t \\ q \\ p \\ \bar{t} \\ \bar{q} \\ \bar{p} \end{pmatrix} = \begin{pmatrix} t+h \\ q + h\nabla_p H(\bar{t},\bar{q},p) \\ p \\ \bar{t} \\ \bar{q} \\ \bar{p} - h\nabla_{\bar{q}} H(\bar{t},\bar{q},p) \end{pmatrix},$$

$$\phi_h^C \begin{pmatrix} t \\ q \\ p \\ \bar{t} \\ \bar{q} \\ \bar{p} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} 2t \\ q+\bar{q} + \cos(2\xi h)(q-\bar{q}) + \sin(2\xi h)(p-\bar{p}) \\ p+\bar{p} - \sin(2\xi h)(q-\bar{q}) + \cos(2\xi h)(p-\bar{p}) \\ 2\bar{t} \\ q+\bar{q} - \cos(2\xi h)(q-\bar{q}) - \sin(2\xi h)(p-\bar{p}) \\ p+\bar{p} + \sin(2\xi h)(q-\bar{q}) - \cos(2\xi h)(p-\bar{p}) \end{pmatrix}.$$

A presymplectic integrator can then be constructed by composing these maps. For instance, with the Strang composition ($r=2$):

$$\phi_{h/2}^A \circ \phi_{h/2}^B \circ \phi_h^C \circ \phi_{h/2}^B \circ \phi_{h/2}^A.$$

Conclusions

- We introduced "presymplectic integrators" which are suitable to simulating dissipative Hamiltonian systems.
- We showed how the important properties of symplectic integrators, which only apply for conservative systems, can be extended to dissipative systems for which there is no underlying conservation law.
- This has implications for optimization; e.g., it allows us to show that presymplectic integrators can yield "rate-matching" optimization algorithms.
- No discrete-time convergence analysis was necessary; it can be guaranteed directly from this framework.
- There is an entire class of algorithms that can be systematically constructed within this framework, and will be guaranteed to preserve the stability and continuous-time rates of convergence.