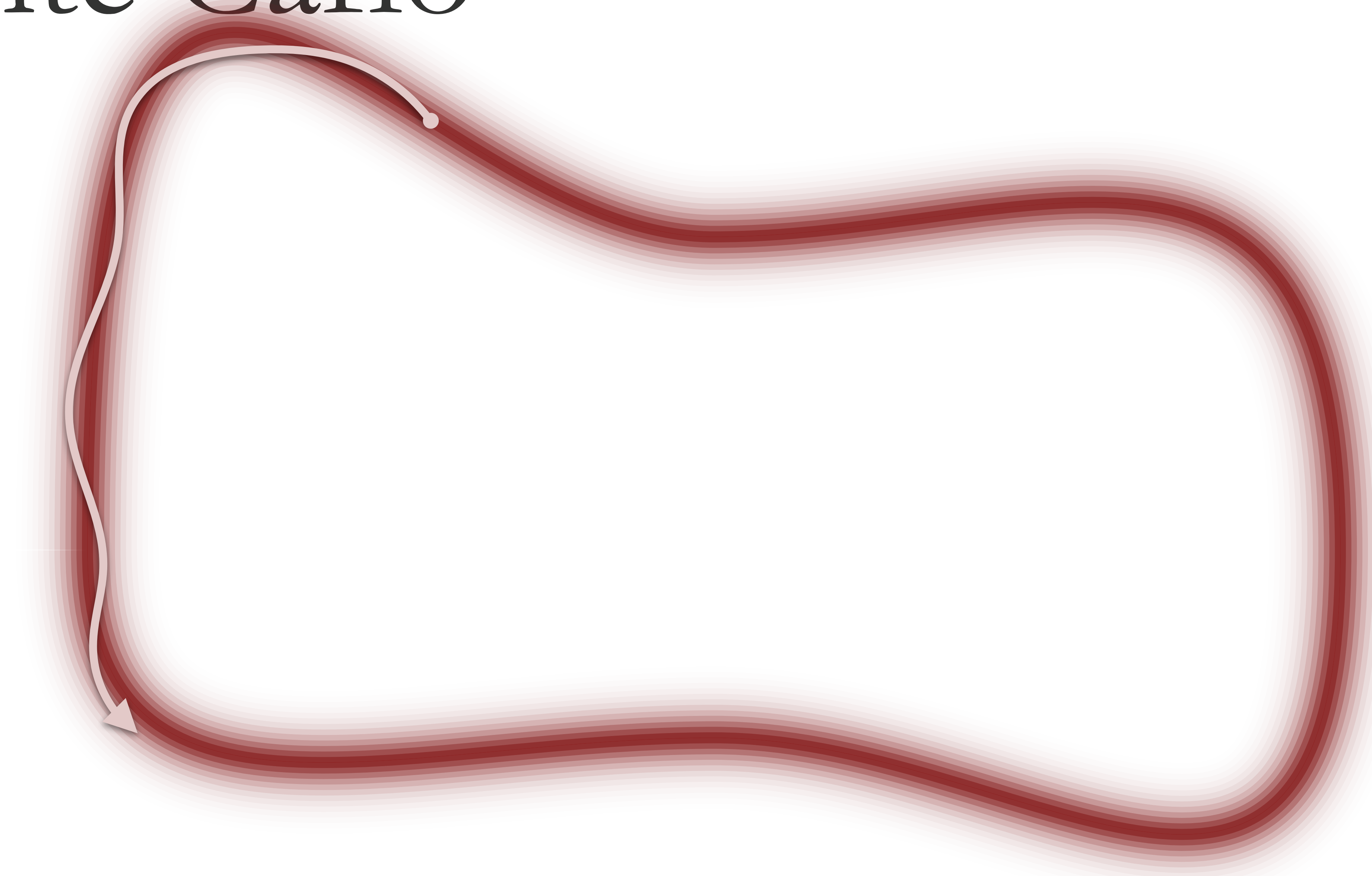
Hamiltonian Monte Carlo

Michael Betancourt
@betanalpha
Symplectomorphic, LLC

Machine Learning Summer School
London, United Kingdom
July 23, 2019

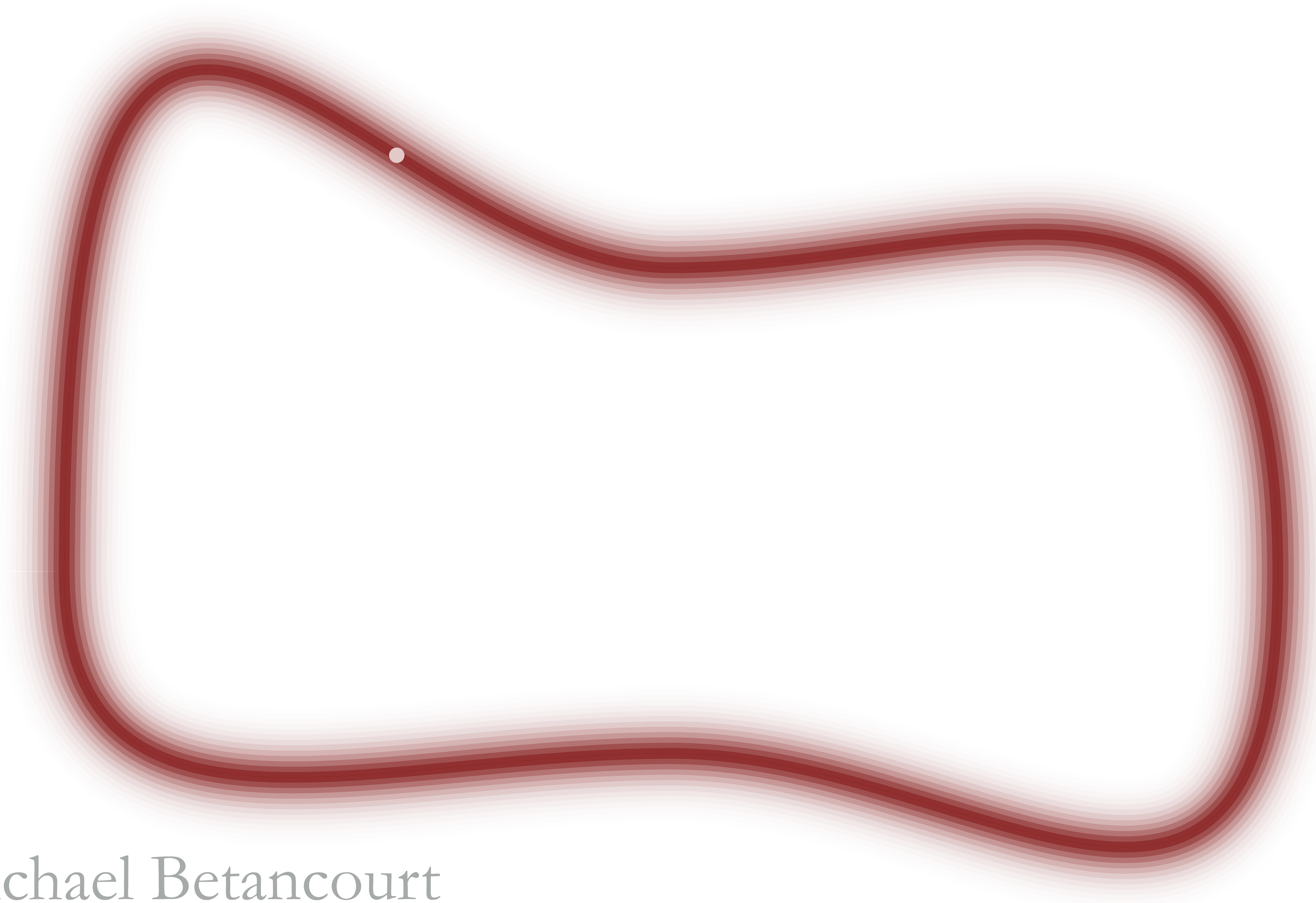
Hamiltonian Monte Carlo

Michael Betancourt
@betanalpha
Symplectomorphic, LLC

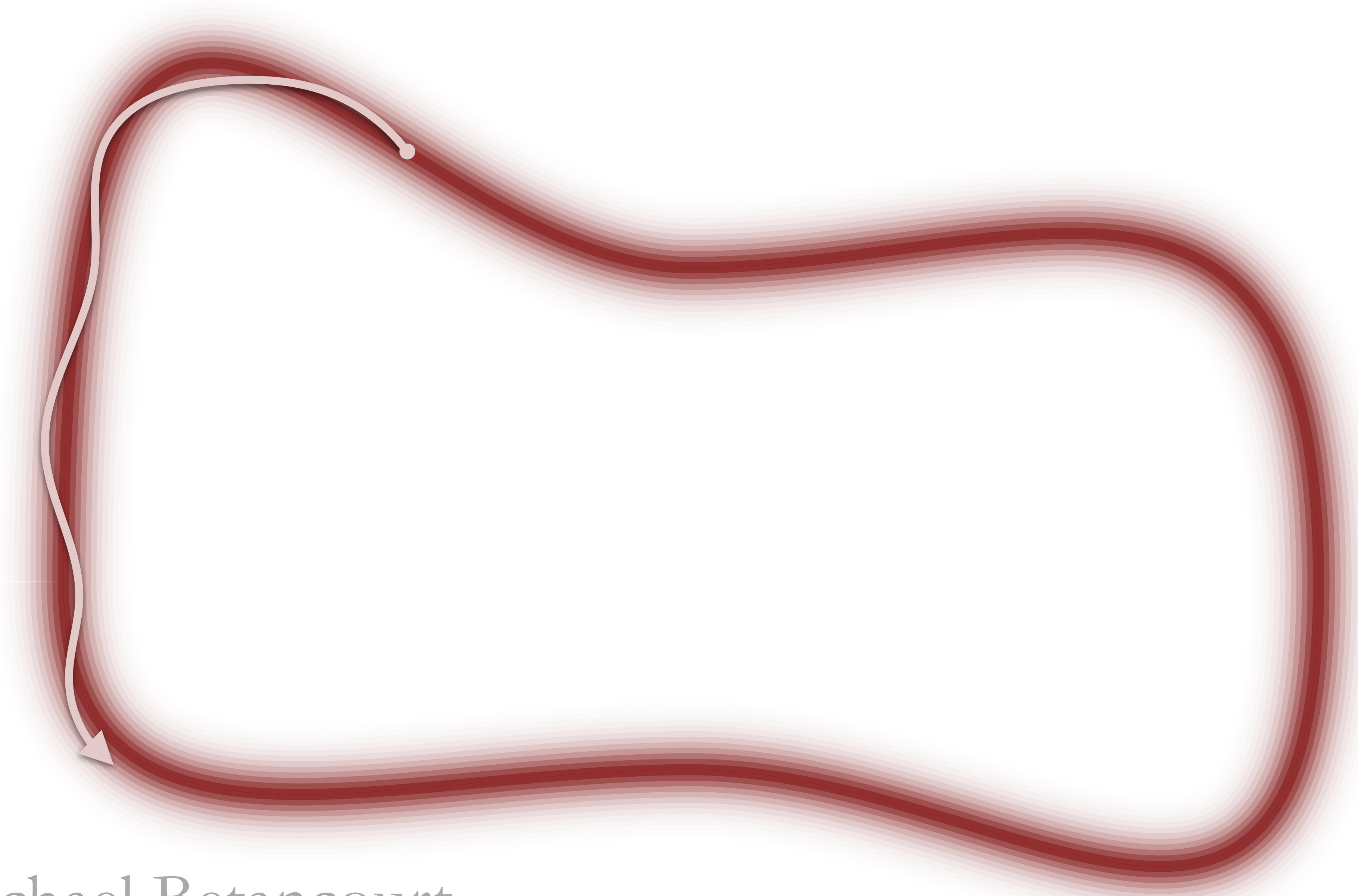


Machine Learning Summer School
London, United Kingdom
July 23, 2019

In order to efficiently scale to complex posteriors, MCMC needs *coherent* exploration of the typical set.



In order to efficiently scale to complex posteriors, MCMC needs *coherent* exploration of the typical set.



Exploration of the typical set is generated by measure-preserving maps over parameter space.

$$\phi_t : Q \to Q$$

$$\left(\pi \circ \phi_t^{-1} \right) (A) = \pi(A)$$

Coherency of that exploration requires equipping a space of maps with the structure of a Lie group to give a *flow*.

$$\phi_t : Q \to Q$$

Coherency of that exploration requires equipping a space of maps with the structure of a Lie group to give a *flow*.

$$\phi_t : Q \to Q$$

$$\phi_t \circ \phi_s = \phi_{s+t}$$

Coherency of that exploration requires equipping a space of maps with the structure of a Lie group to give a *flow*.

$$\phi_t : Q \to Q$$

$$\phi_t \circ \phi_s = \phi_{s+t}$$

$$\phi_t^{-1} = \phi_{-t}$$

$$\phi_0 = \mathrm{Id}_Q$$

Together these conditions yield *measure-preserving flows* which coherently explore the typical set as desired.

$$\phi_t : Q \to Q$$

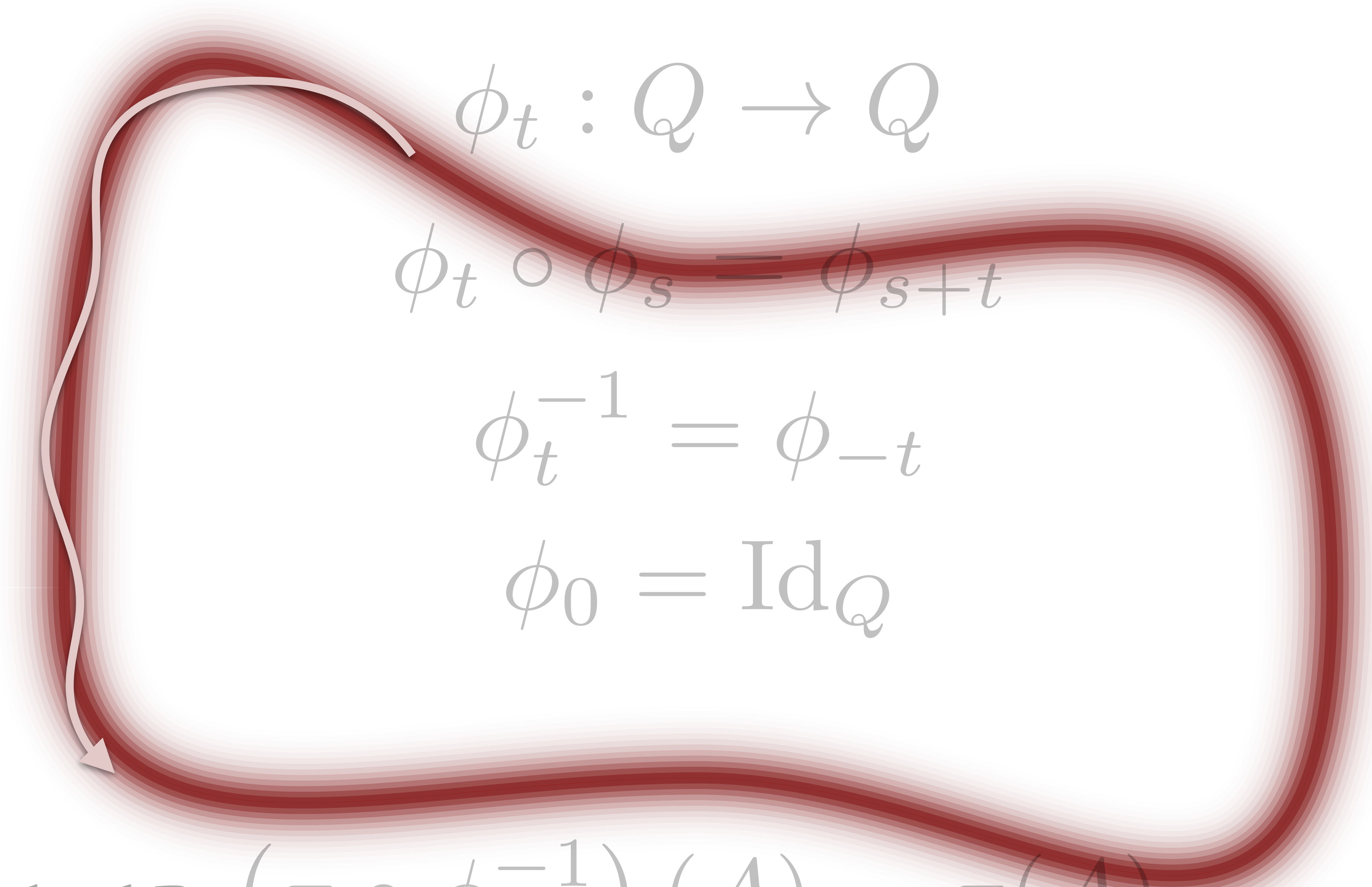
$$\phi_t \circ \phi_s = \phi_{s+t}$$

$$\phi_t^{-1} = \phi_{-t}$$

$$\phi_0 = \mathrm{Id}_Q$$

$$\left(\pi \circ \phi_t^{-1}\right)(A) = \pi(A)$$

Together these conditions yield *measure-preserving flows* which coherently explore the typical set as desired.



$$\phi_t : Q \to Q$$

$$\phi_t \circ \phi_s = \phi_{s+t}$$

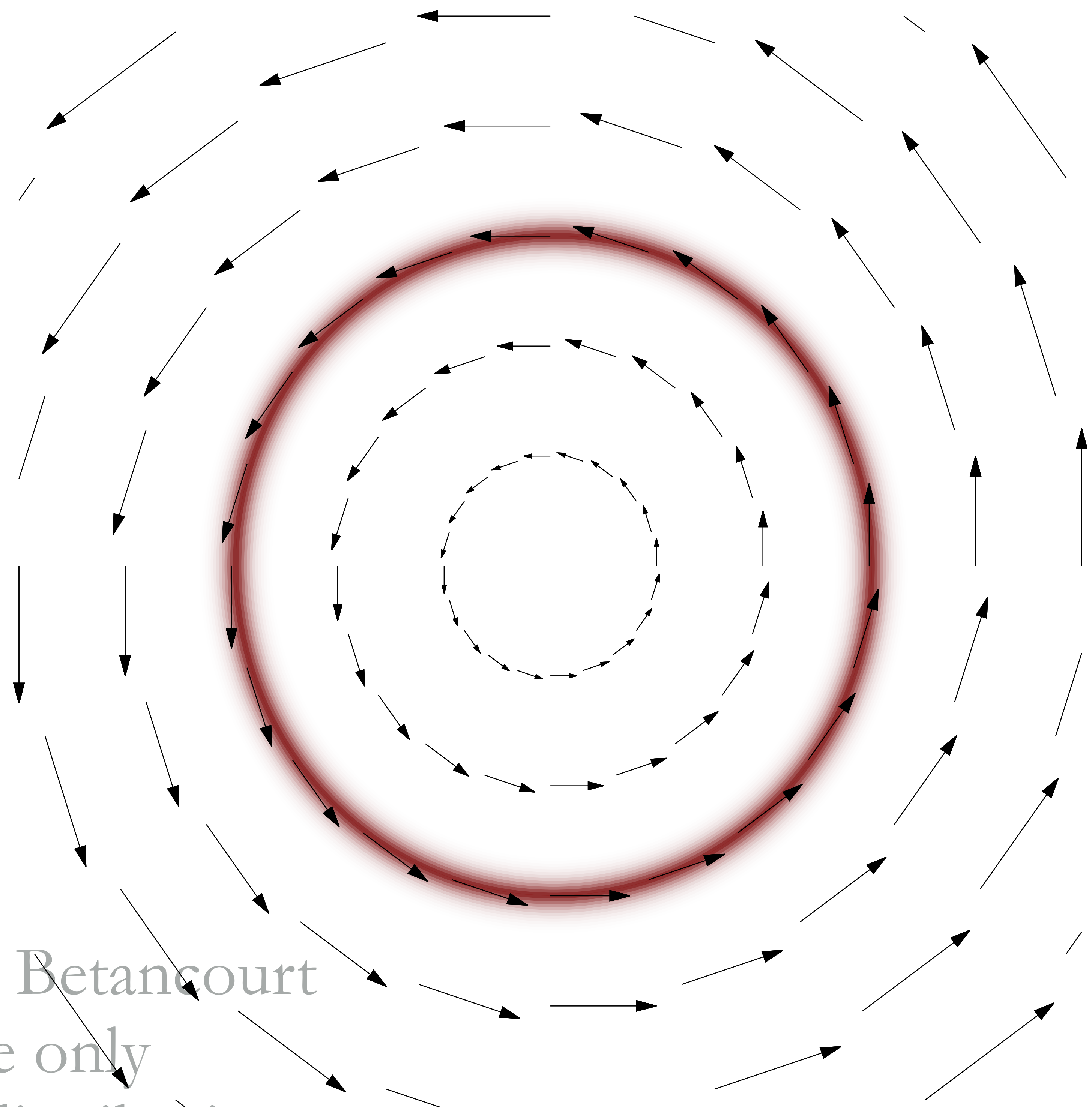
$$\phi_t^{-1} = \phi_{-t}$$

$$\phi_0 = \mathrm{Id}_Q$$

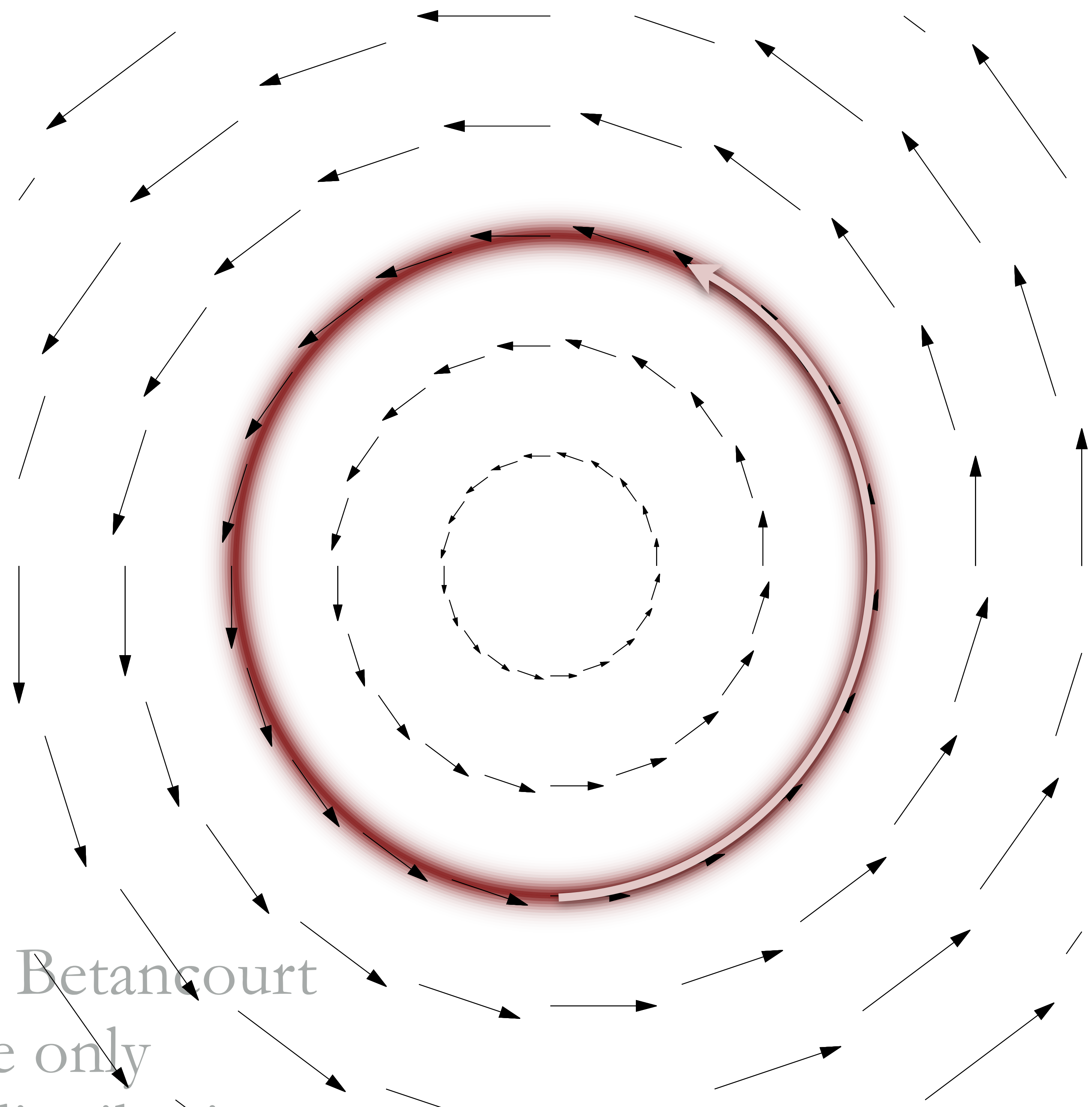
$$\left(\pi \circ \phi_t^{-1}\right)(A) = \pi(A)$$

One way to construct the desired exploration is to integrate along a *vector field* aligned with the typical set.

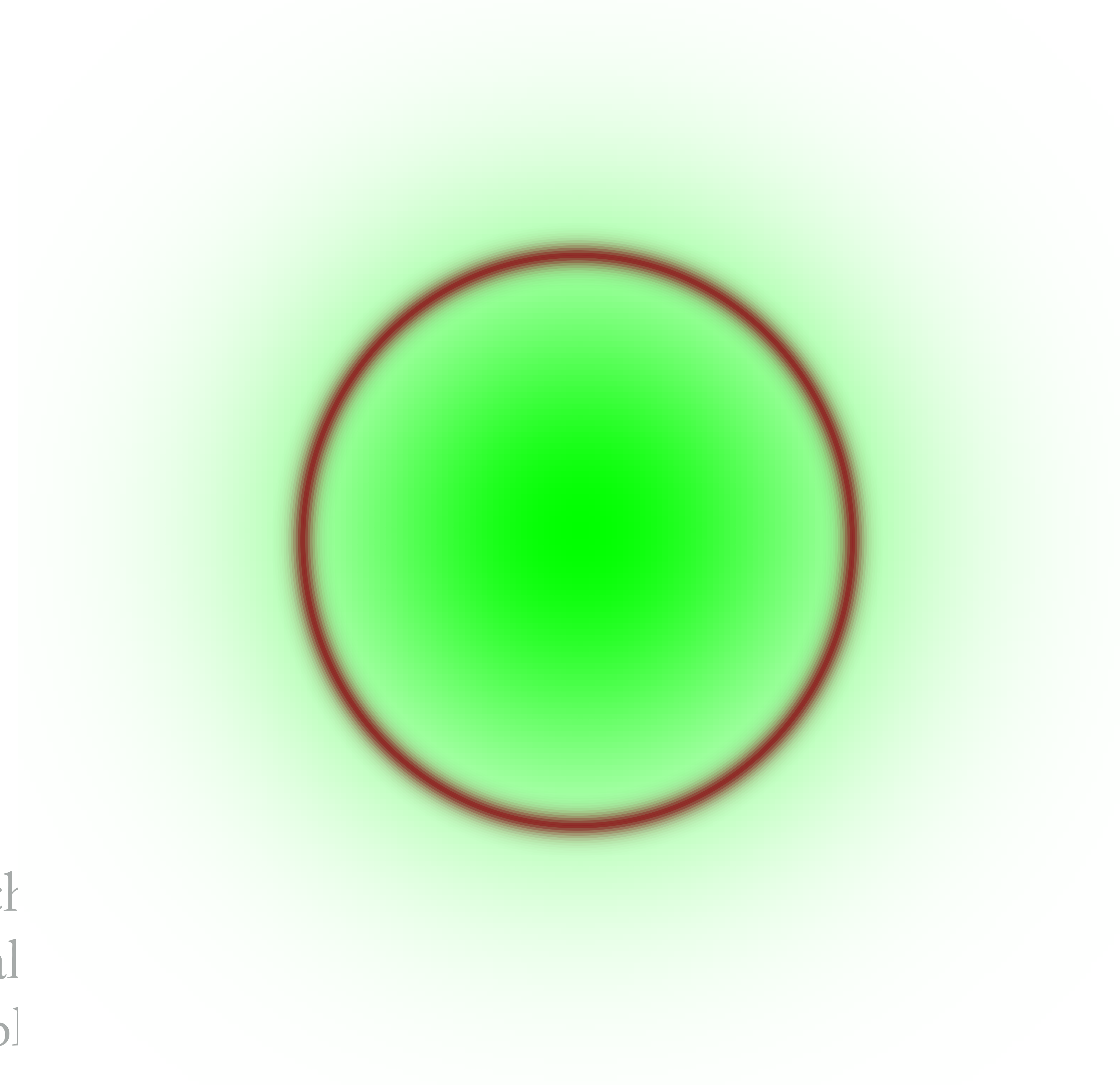
One way to construct the desired exploration is to integrate along a *vector field* aligned with the typical set.



One way to construct the desired exploration is to integrate along a *vector field* aligned with the typical set.

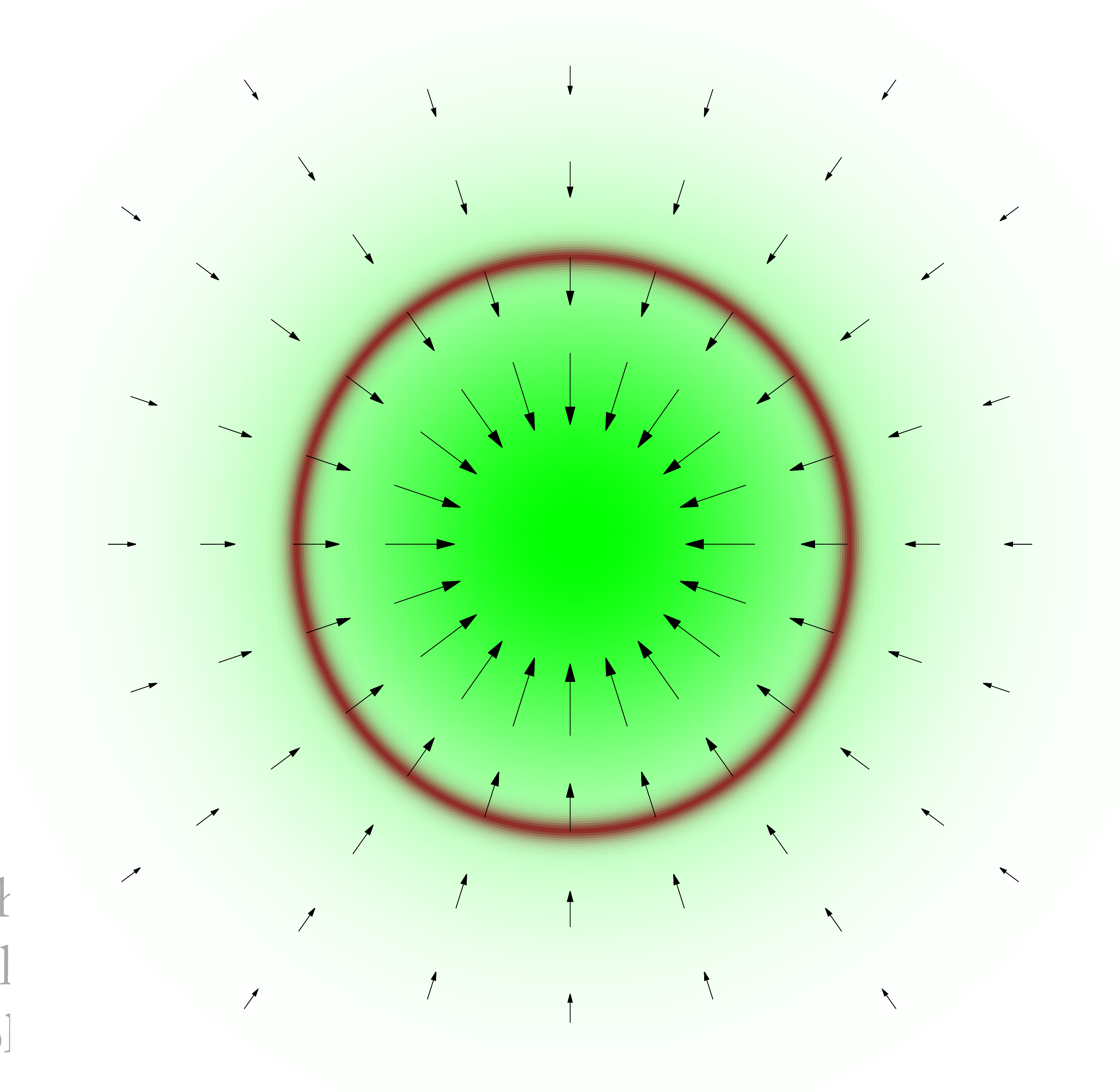


Creating the desired vector field requires transforming available vector fields, such as the gradient.



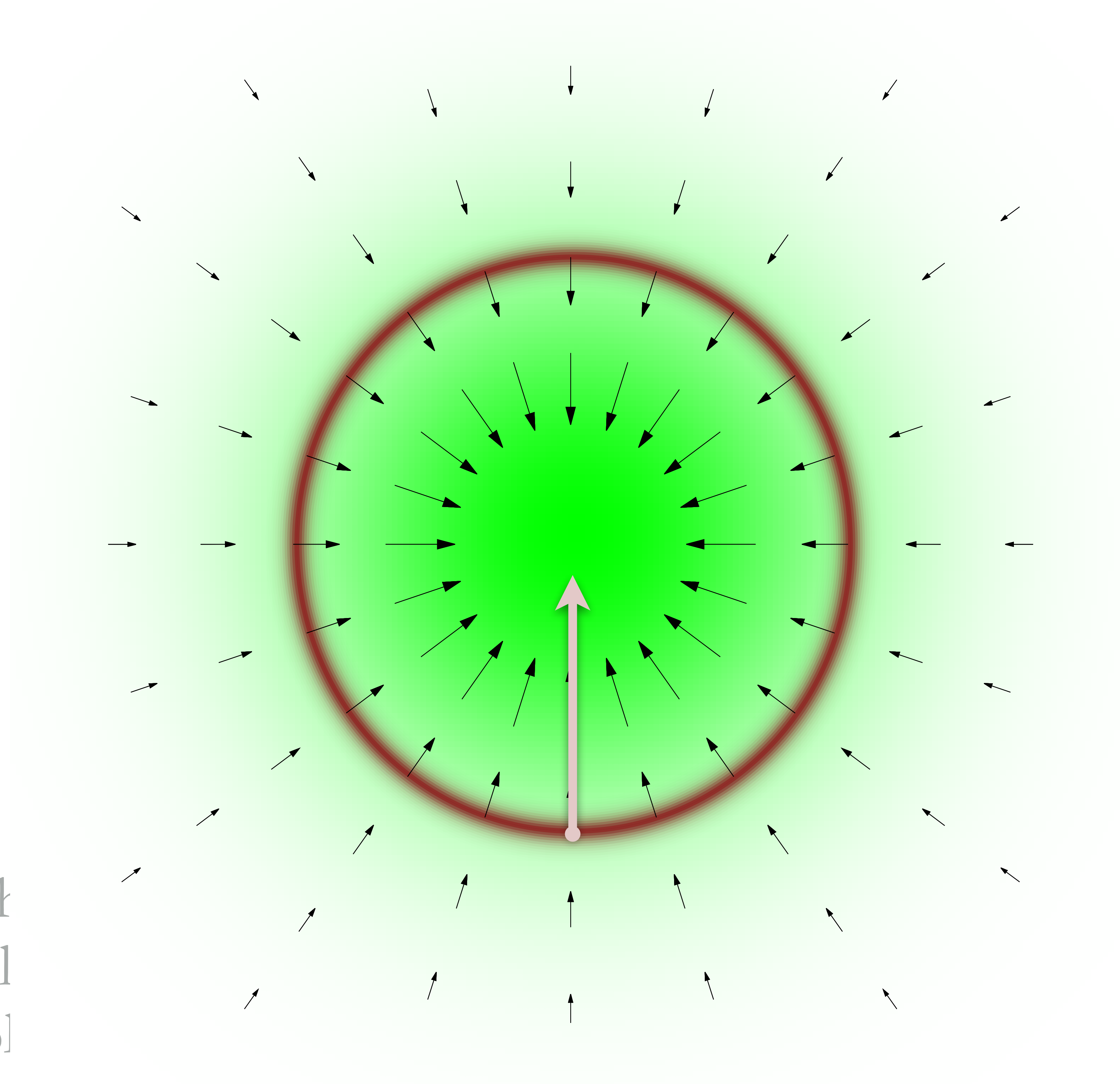
$\pi(q)$

Creating the desired vector field requires transforming available vector fields, such as the gradient.



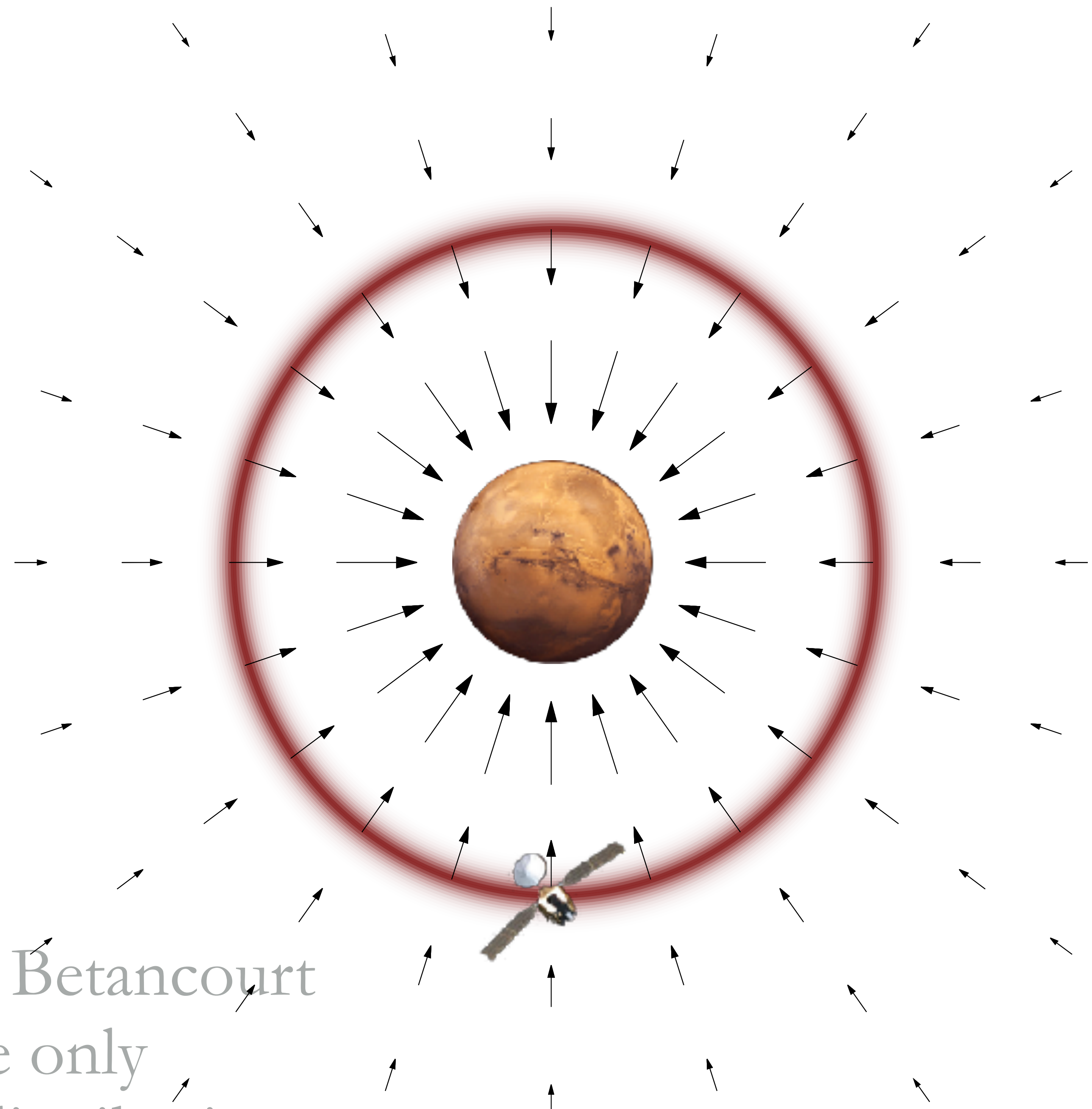
$$\frac{\partial \pi(q)}{\partial q}$$

Creating the desired vector field requires transforming available vector fields, such as the gradient.

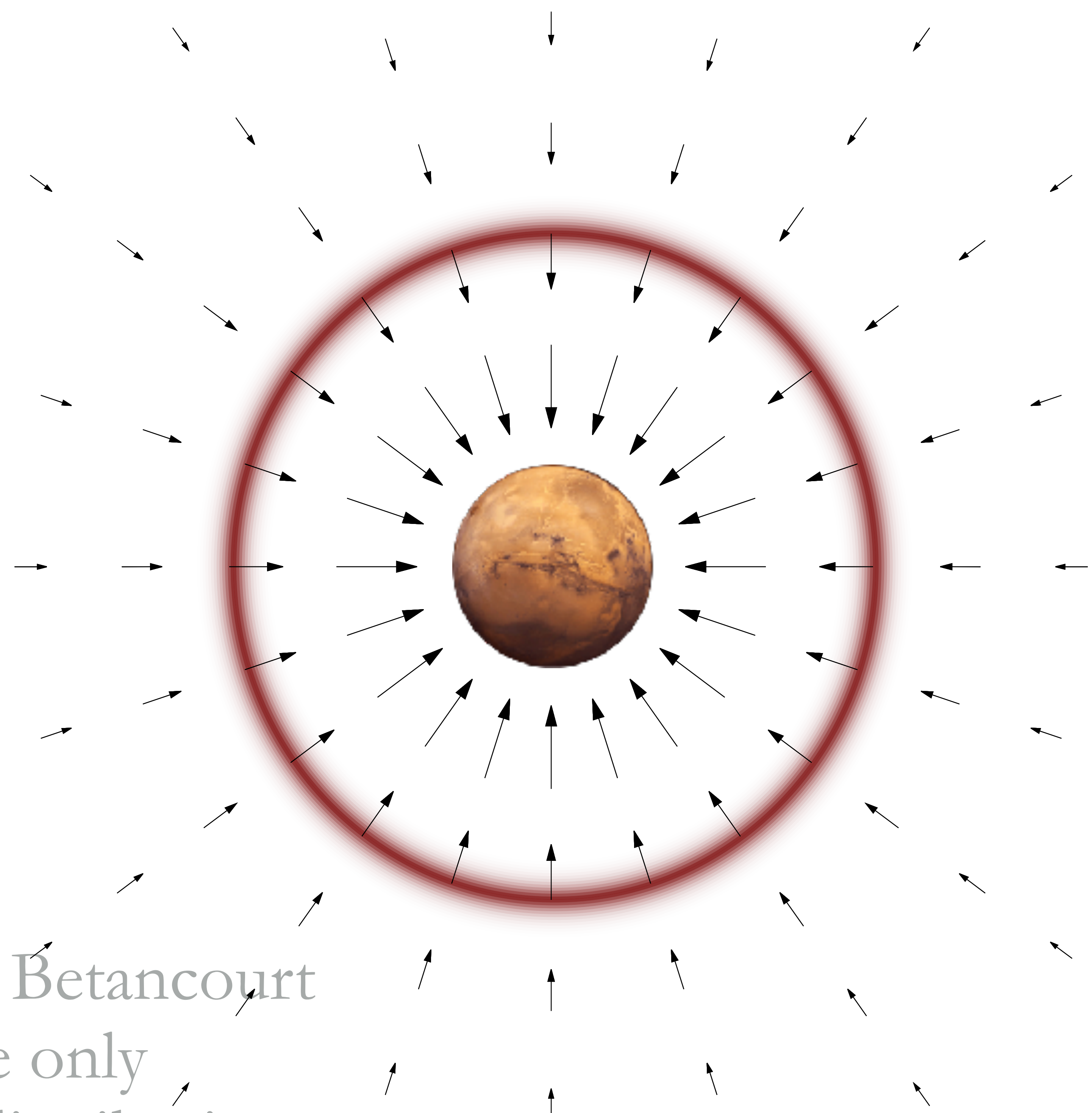


$$\frac{\partial \pi(q)}{\partial q}$$

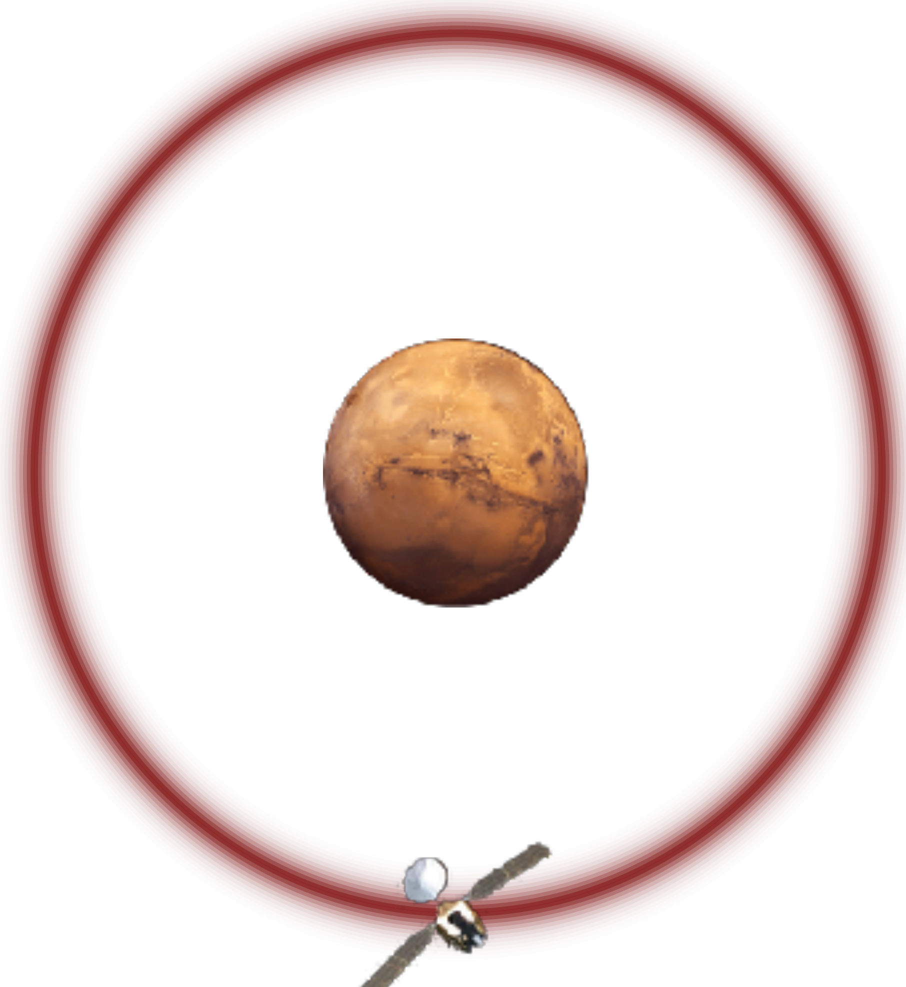
Differential geometry informs this transformation, although a physical analogy can be more intuitive.



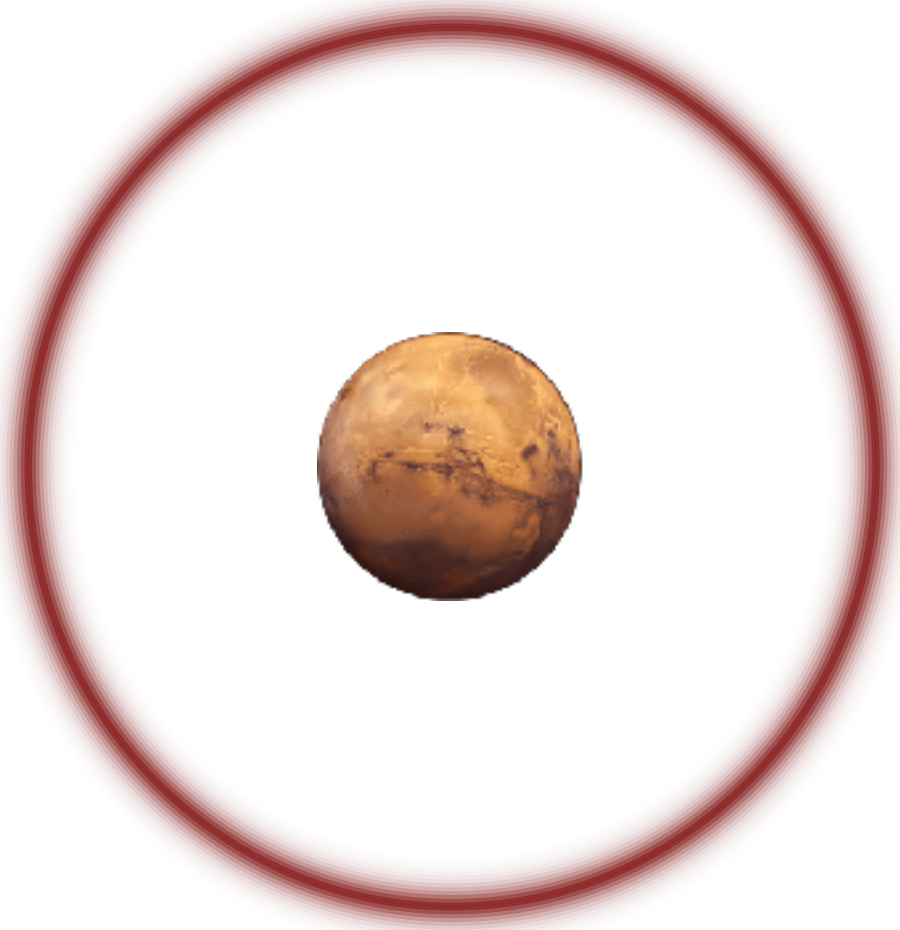
Differential geometry informs this transformation, although a physical analogy can be more intuitive.



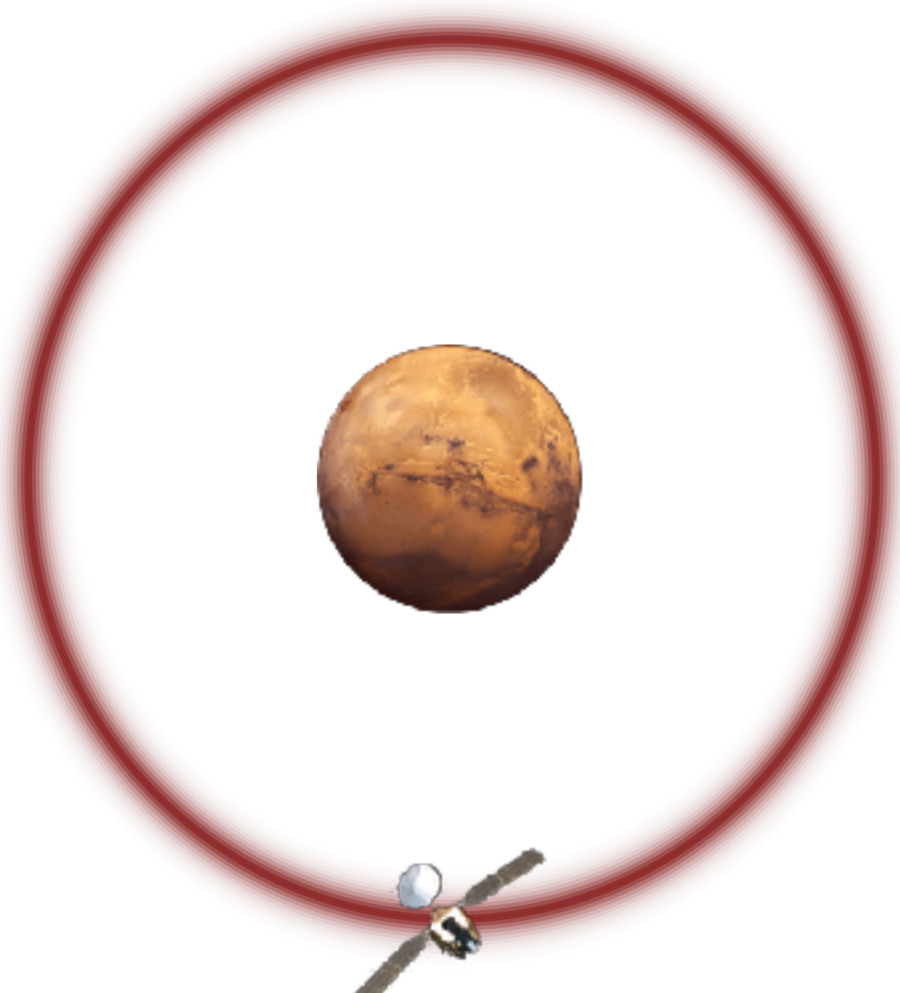
We need to add *momentum* in just the right way.
Too little and we still crash into the planet.



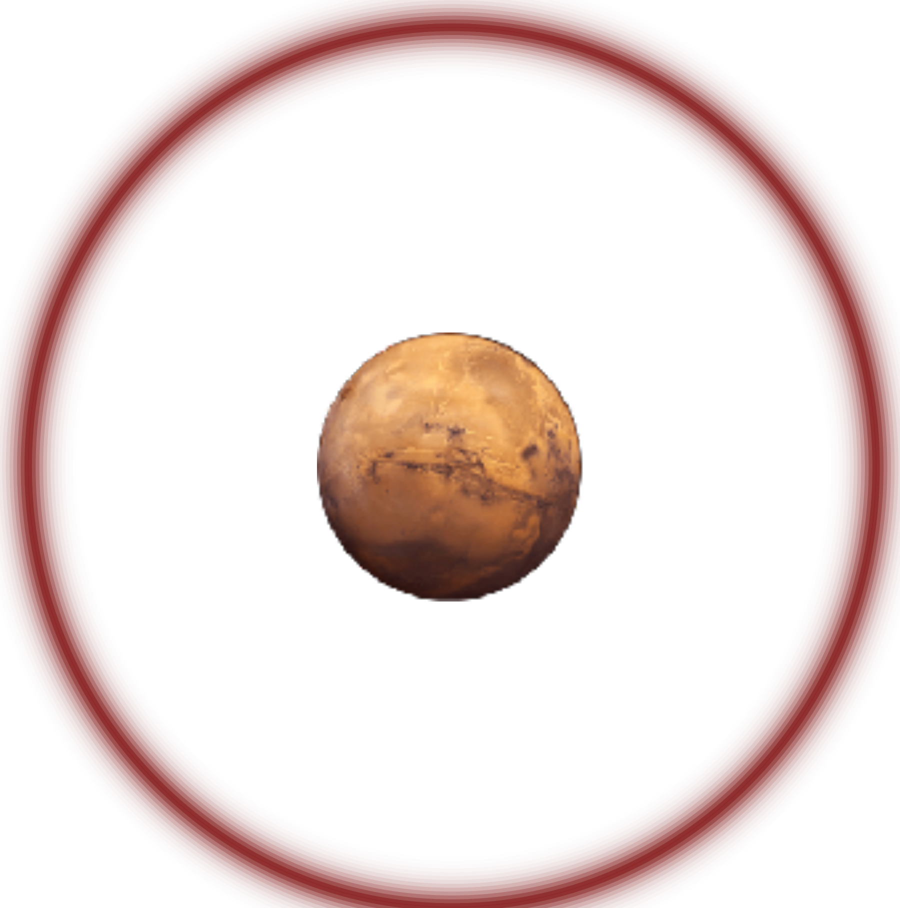
We need to add *momentum* in just the right way.
Too little and we still crash into the planet.



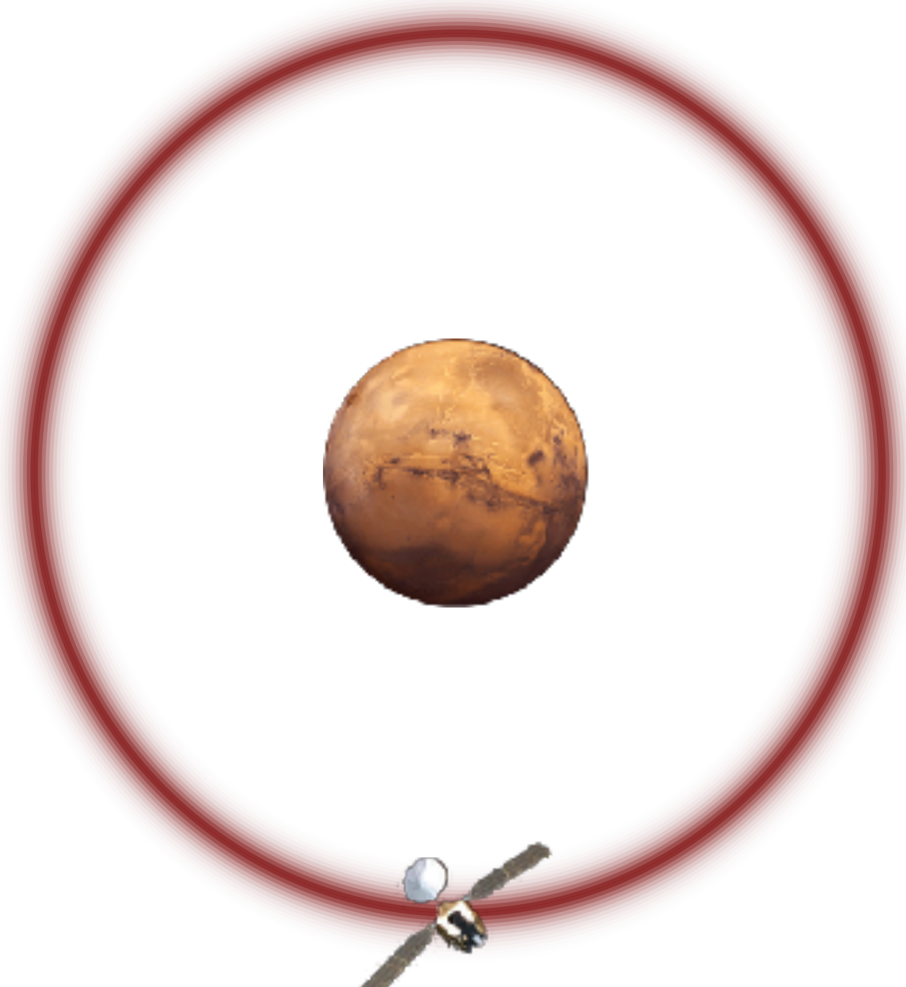
Too much and we fly off to infinity.



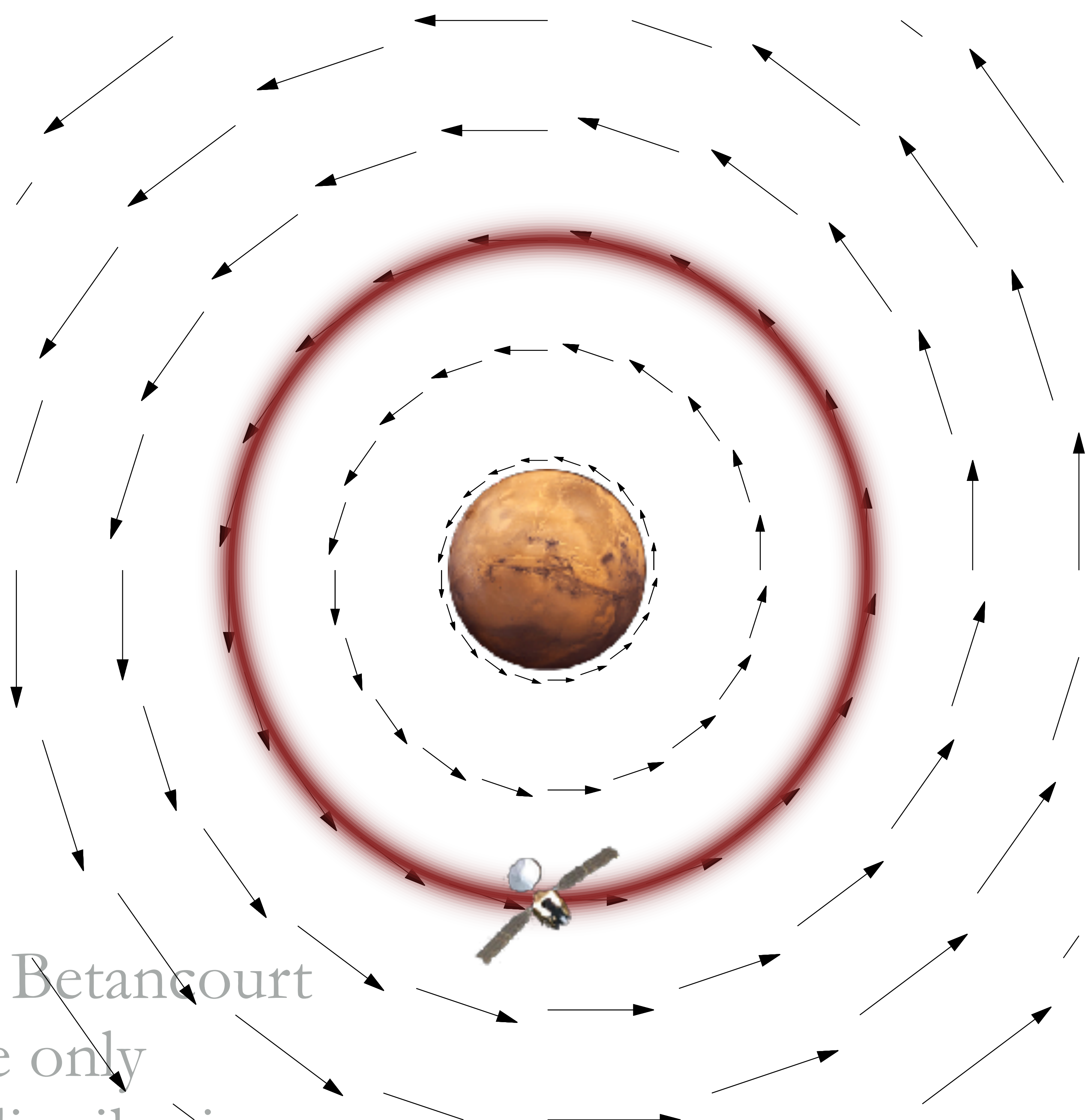
Too much and we fly off to infinity.



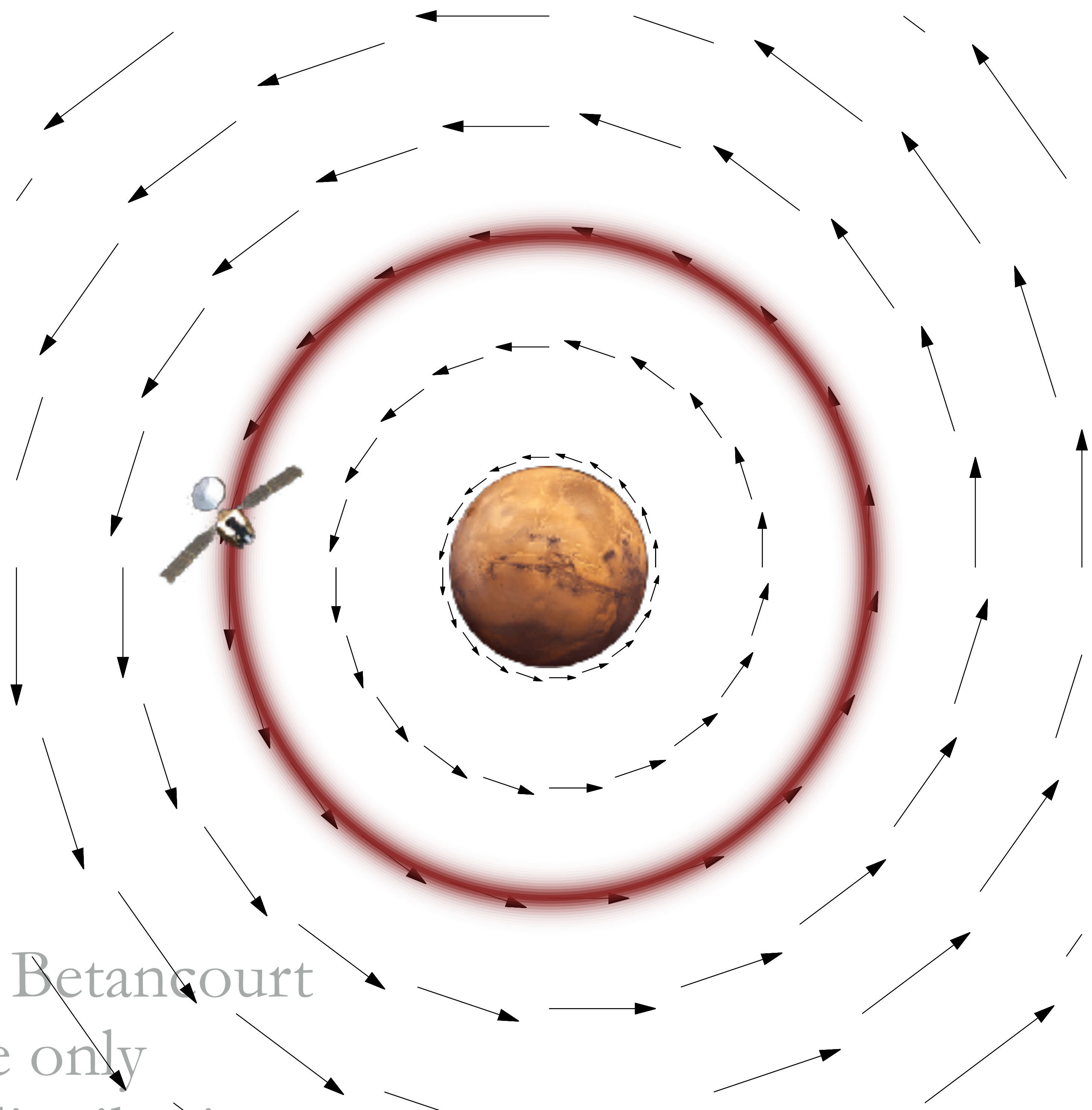
Just enough, however, aligns the gradients with the typical set and yields the desired orbital trajectory.



Just enough, however, aligns the gradients with the typical set and yields the desired orbital trajectory.



Just enough, however, aligns the gradients with the typical set and yields the desired orbital trajectory.



Hamiltonian Monte Carlo is the formal procedure for adding momentum to generate coherent exploration.

$$q \to (q, p)$$

Hamiltonian Monte Carlo is the formal procedure for adding momentum to generate coherent exploration.

$$q \to (q, p)$$

$$\pi(q) \to \pi(q, p)$$

Hamiltonian Monte Carlo is the formal procedure for adding momentum to generate coherent exploration.

$$q \to (q, p)$$

$$\pi(q) \to \pi(q, p) = \pi(p \mid q)\, \pi(q)$$

Lifting the target distribution gives a *Hamiltonian* that decomposes into a *potential energy* and a *kinetic energy*.

$$\pi(p \mid q) \, \pi(q)$$

Lifting the target distribution gives a *Hamiltonian* that decomposes into a *potential energy* and a *kinetic energy*.

$$H(q,p) = -\log \pi(p \mid q)\,\pi(q)$$

Lifting the target distribution gives a *Hamiltonian* that decomposes into a *potential energy* and a *kinetic energy*.

$$H(q,p) = -\log \pi(p \mid q)\, \pi(q)$$

$$= -\log \pi(p \mid q) - \log \pi(q)$$

Lifting the target distribution gives a *Hamiltonian* that decomposes into a *potential energy* and a *kinetic energy*.

$$H(q,p) = -\log \pi(p \mid q)\, \pi(q)$$

$$= \underbrace{-\log \pi(p \mid q)}_{K} - \log \pi(q)$$

Lifting the target distribution gives a *Hamiltonian* that decomposes into a *potential energy* and a *kinetic energy*.

$$H(q,p) = -\log \pi(p \mid q) \, \pi(q)$$

$$= \underbrace{-\log \pi(p \mid q)}_{K} \underbrace{- \log \pi(q)}_{V}$$

Any choice of kinetic energy generates coherent exploration through the expanded phase space.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial K}{\partial p}$$

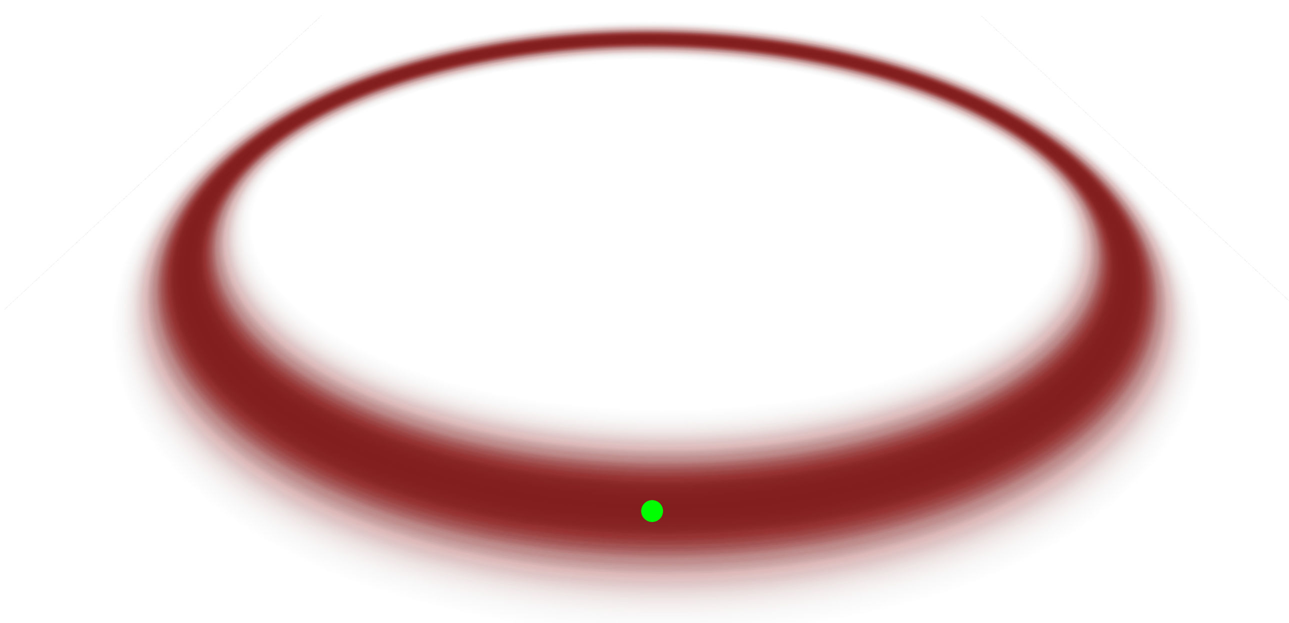
$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}$$

Any choice of kinetic energy generates coherent exploration through the expanded phase space.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial K}{\partial p}$$

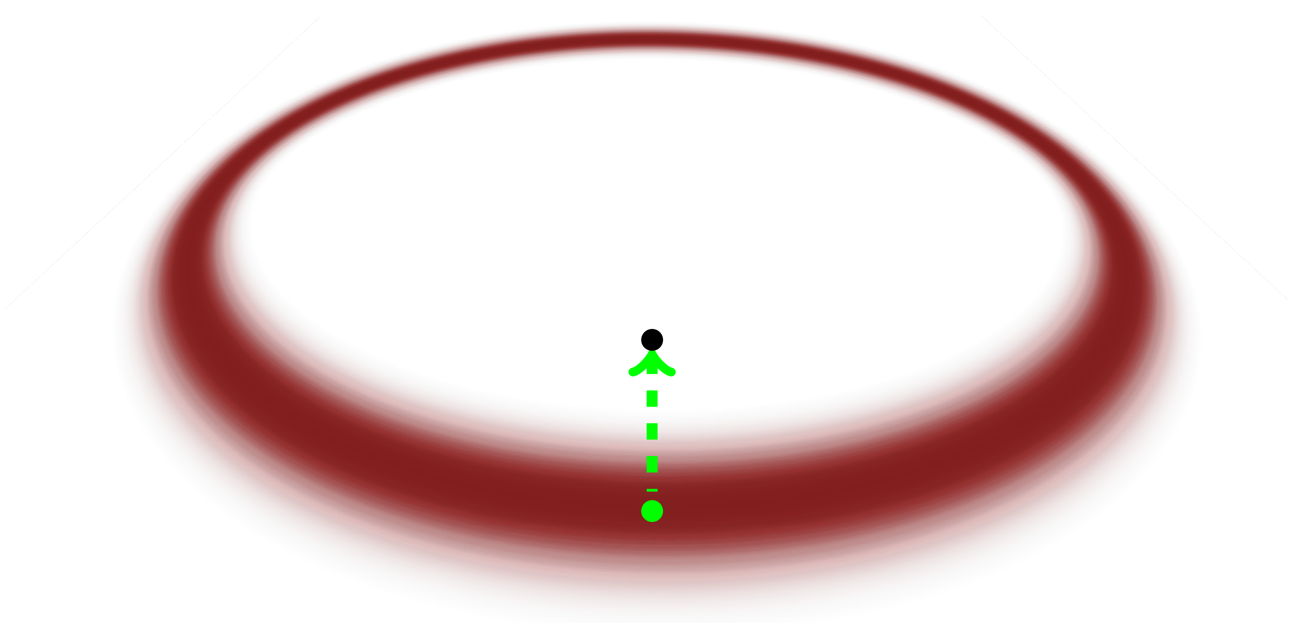
$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}$$

We then construct a Markov transition by lifting into, exploring, and projecting from the expanded space.

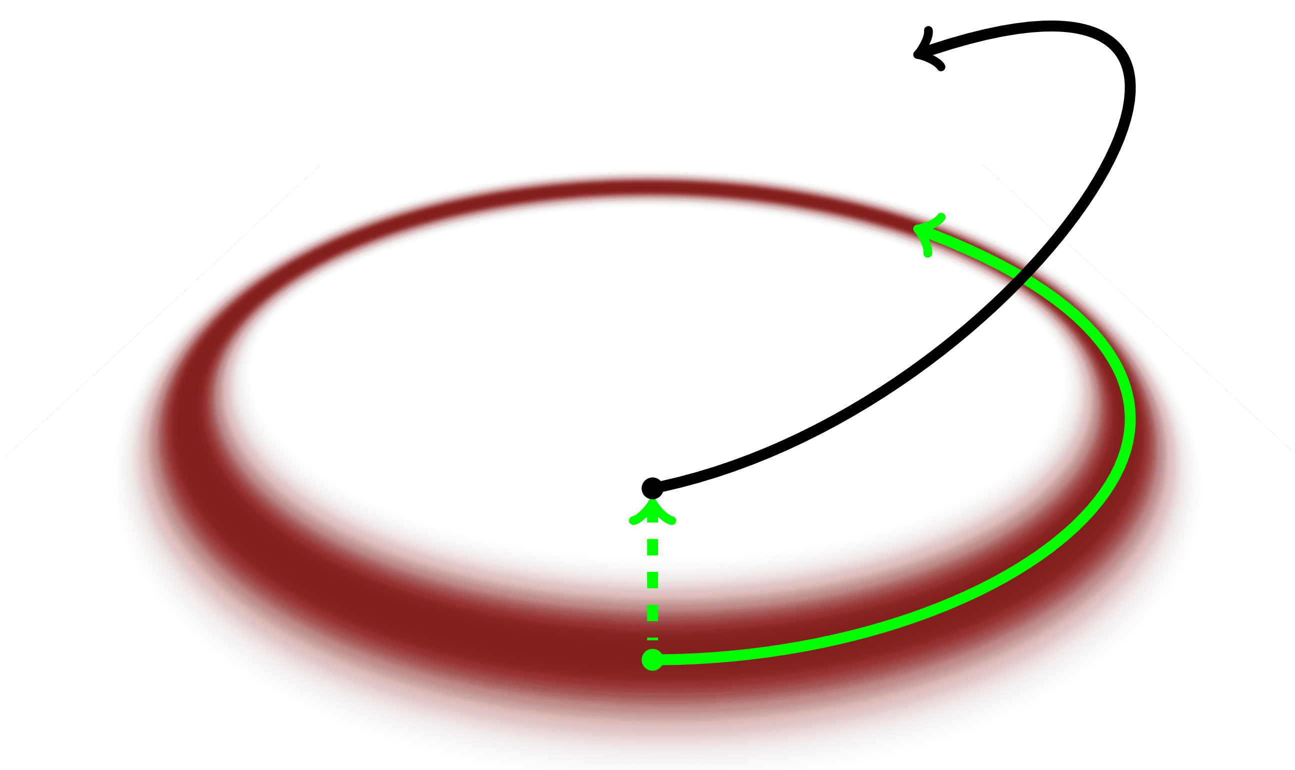


We then construct a Markov transition by lifting into, exploring, and projecting from the expanded space.

$$p \sim \pi(p \mid q)$$



We then construct a Markov transition by lifting into, exploring, and projecting from the expanded space.

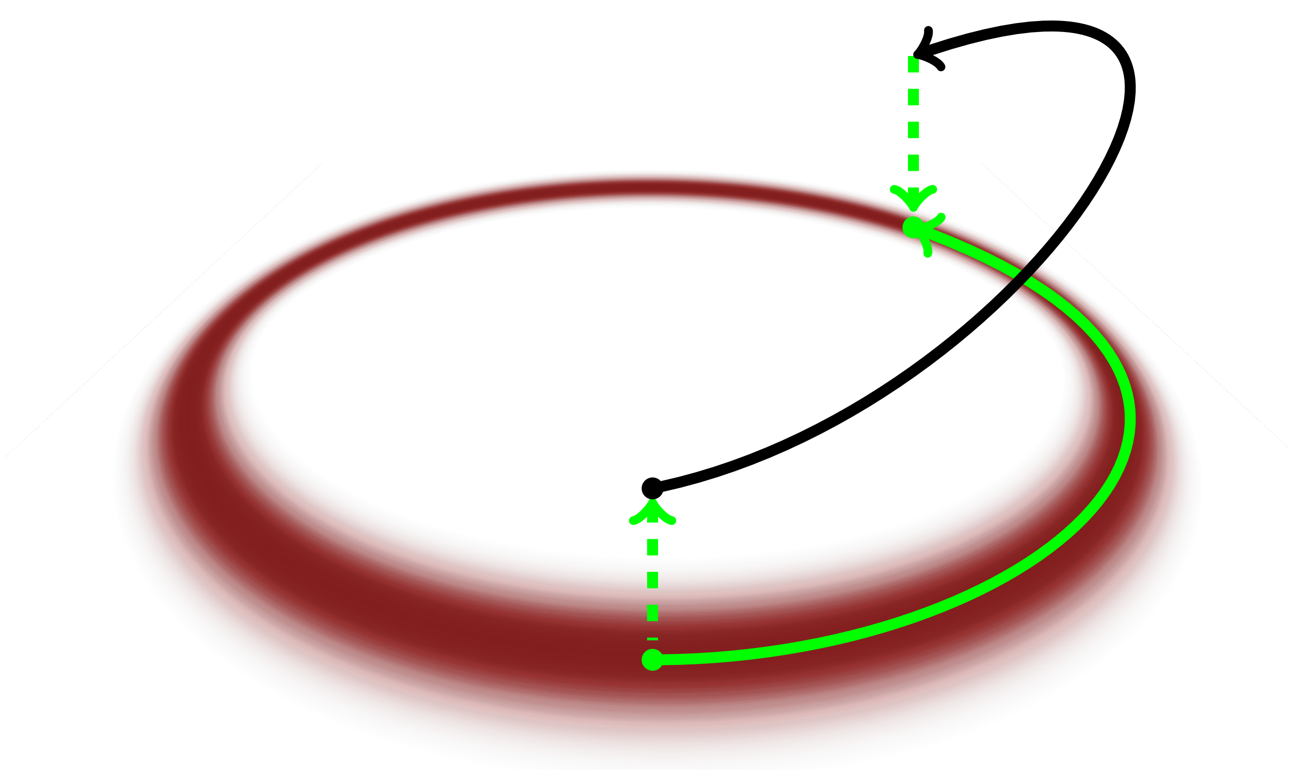


$$p \sim \pi(p \mid q)$$

$$t \sim U[0, T]$$

$$(q', p') \rightarrow \phi_t(q, p)$$

We then construct a Markov transition by lifting into, exploring, and projecting from the expanded space.



$$p \sim \pi(p \mid q)$$

$$t \sim U[0, T]$$

$$(q', p') \to \phi_t(q, p)$$

$$(q', p') \to q'$$

In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

$$\frac{\mathrm{d}q}{\mathrm{d}t} = \frac{\partial K}{\partial p}$$

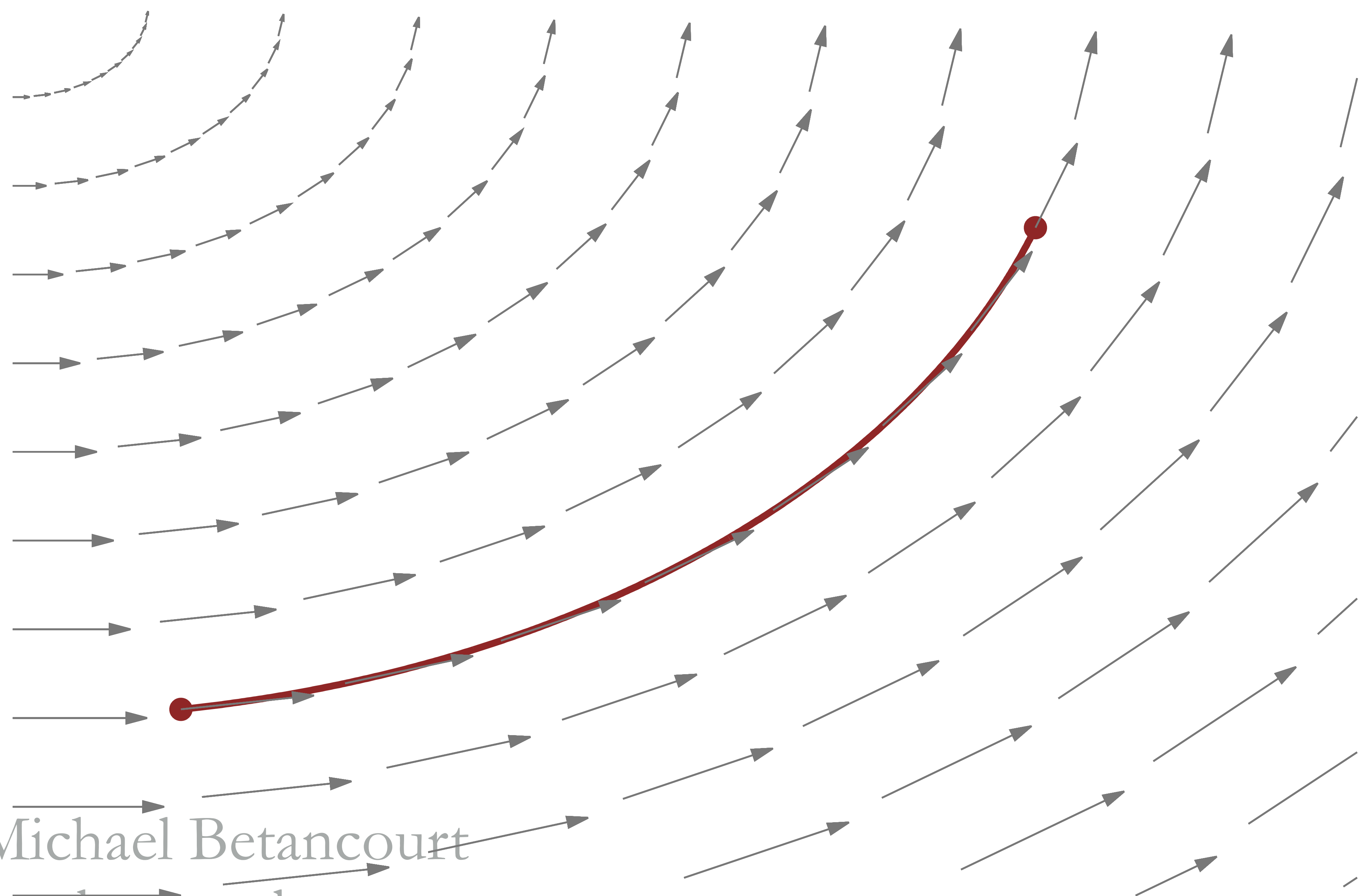
$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial K}{\partial q} - \frac{\partial V}{\partial q}$$

In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

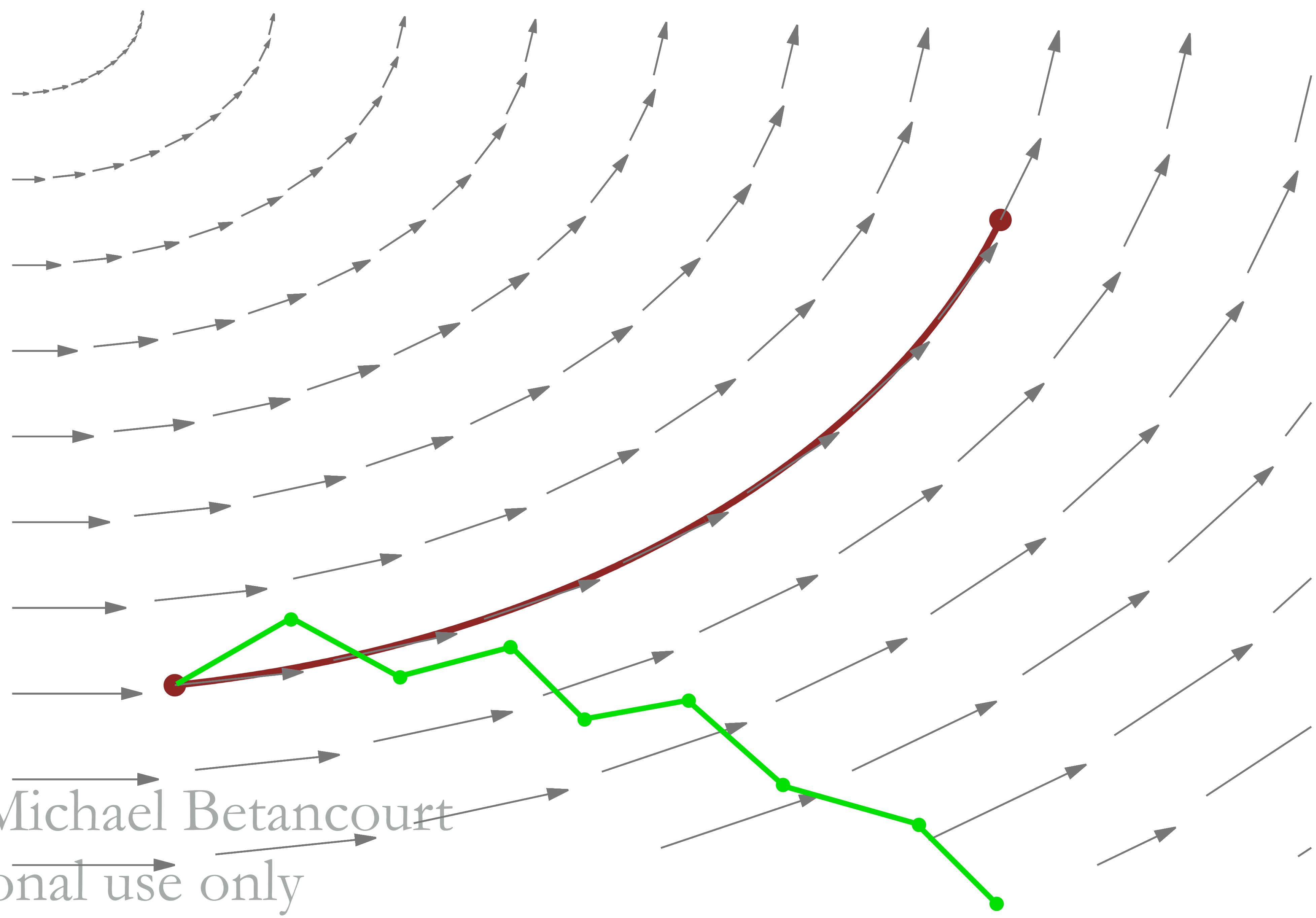
$$q \to q + \epsilon \frac{\partial K}{\partial p}$$

$$p \to p - \epsilon \left(\frac{\partial K}{\partial q} + \frac{\partial V}{\partial q} \right)$$

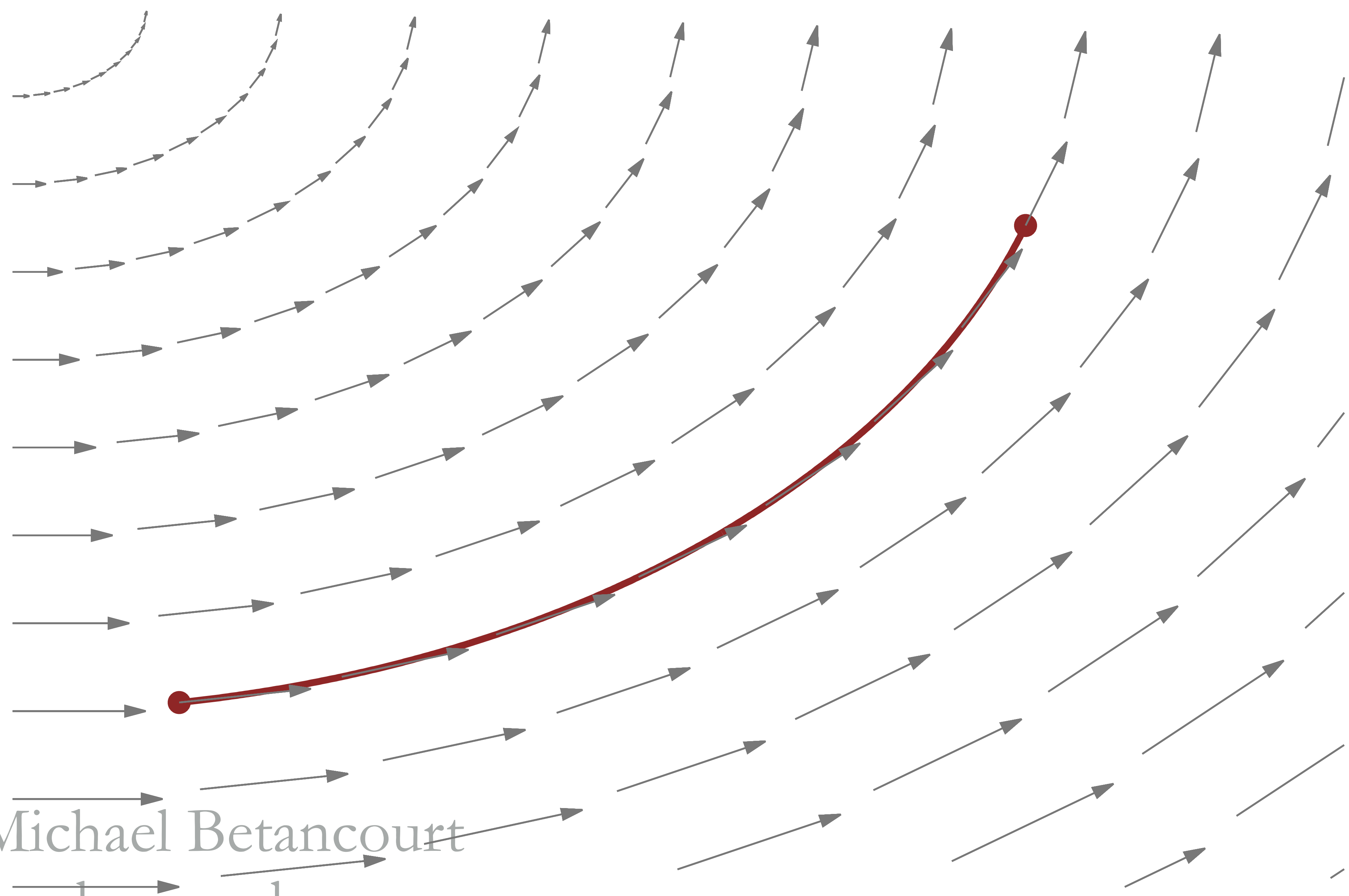
Naive numerical integrators drift away from the exact trajectory, limiting the time for which we can explore.



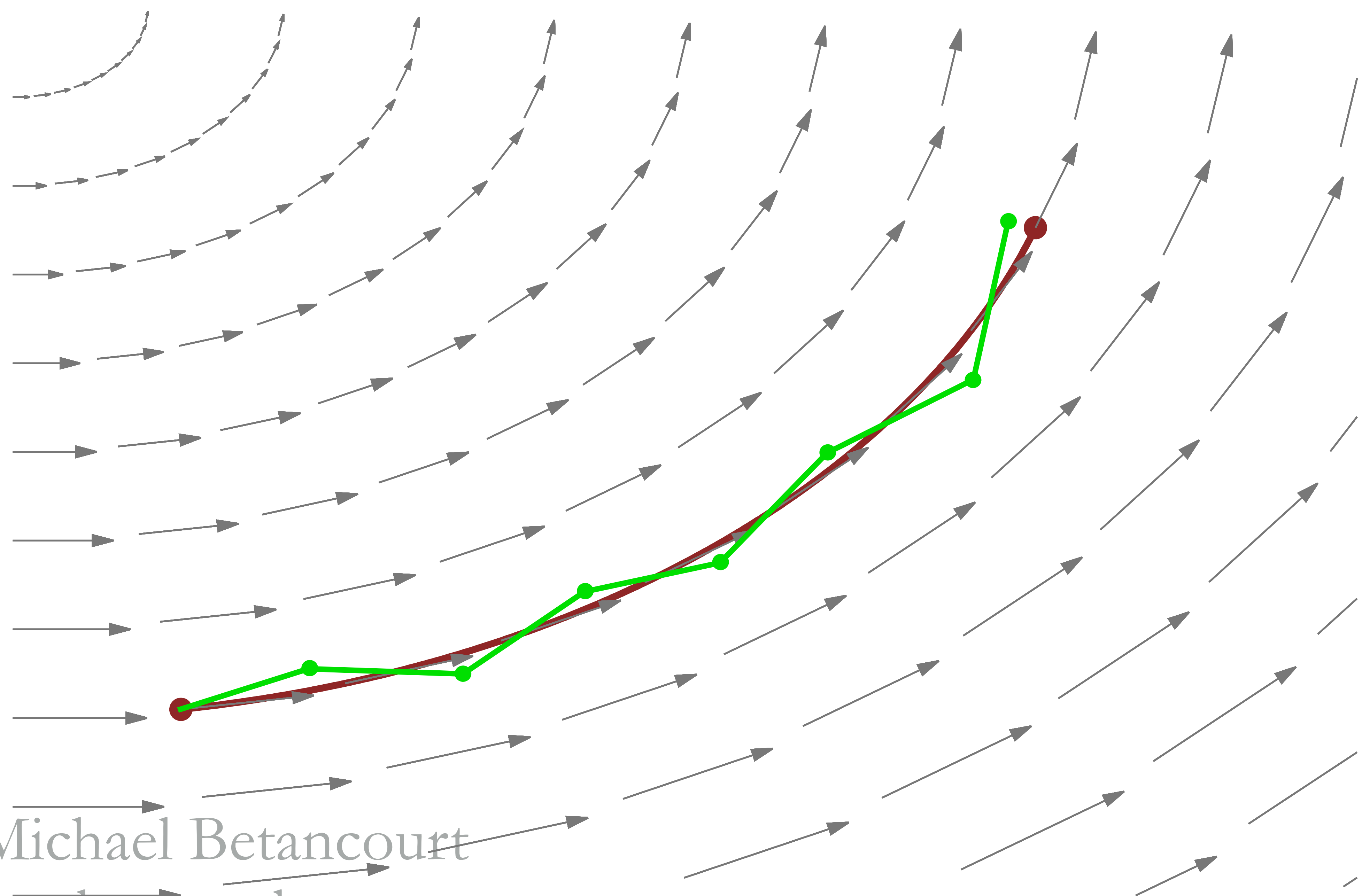
Naive numerical integrators drift away from the exact trajectory, limiting the time for which we can explore.



Symplectic integrators very accurately simulate these Hamiltonian systems, even in high dimensions.



Symplectic integrators very accurately simulate these Hamiltonian systems, even in high dimensions.



In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

$$p \sim \pi(p \mid q)$$

$$t \sim U[0, T]$$

$$(q', p') \to \phi_t(q, p)$$

$$(q', p') \to q'$$

In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

$$p \sim \pi(p \mid q) \qquad \longrightarrow \qquad p \sim \pi(p \mid q)$$

$$t \sim U[0, T]$$

$$(q', p') \to \phi_t(q, p)$$

$$(q', p') \to q'$$

In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

$$p \sim \pi(p \mid q) \quad \longrightarrow \quad p \sim \pi(p \mid q)$$

$$t \sim U[0, T] \quad \longrightarrow \quad L \sim U\left[0, \ldots, T/\epsilon\right]$$

$$(q', p') \to \phi_t(q, p)$$

$$(q', p') \to q'$$

In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

$$p \sim \pi(p \mid q) \quad \longrightarrow \quad p \sim \pi(p \mid q)$$

$$t \sim U[0, T] \quad \longrightarrow \quad L \sim U\left[0, \ldots, T/\epsilon\right]$$

$$(q', p') \to \phi_t(q, p) \quad \longrightarrow \quad (q', p') \to \Phi_{\epsilon, L}(q, p)$$

$$(q', p') \to q'$$

In practice we have to approximate the exact Hamiltonian trajectories with numerical trajectories.

$$p \sim \pi(p \mid q) \quad \longrightarrow \quad p \sim \pi(p \mid q)$$

$$t \sim U[0, T] \quad \longrightarrow \quad L \sim U\left[0, \ldots, T/\epsilon\right]$$

$$(q', p') \to \phi_t(q, p) \quad \longrightarrow \quad (q', p') \to \Phi_{\epsilon, L}(q, p)$$

$$(q', p') \to q' \quad \longrightarrow \quad (q', p') \to q'$$

These numerical trajectories introduce error that we need to correct for lest we bias our Markov chain.

$$p \sim \pi(p \mid q)$$

$$L \sim U\left[0, \ldots, T/\epsilon\right]$$

$$(q', p') \to \Phi_{\epsilon, L}(q, p)$$

$$(q', p') \to q'$$

These numerical trajectories introduce error that we need to correct for lest we bias our Markov chain.

$$p \sim \pi(p \mid q)$$

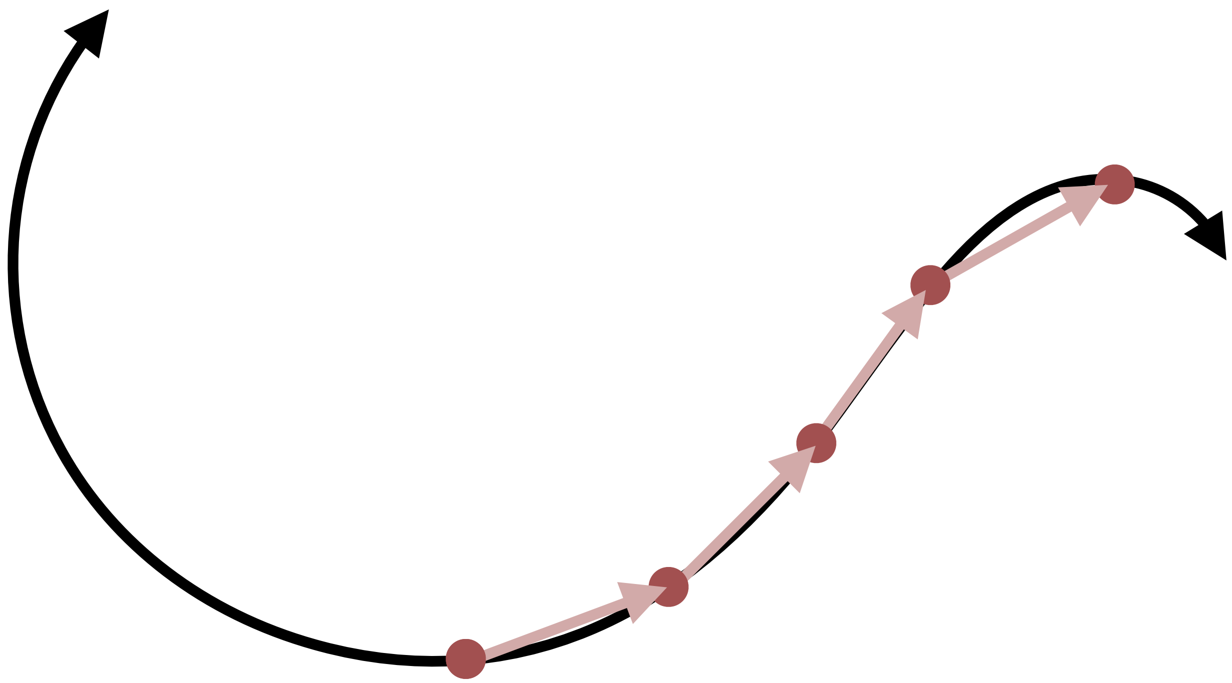
$$L \sim U\left[0, \ldots, T/\epsilon\right]$$

$$(q', p') \to \Phi_{\epsilon, L}(q, p)$$

?

$$(q', p') \to q'$$

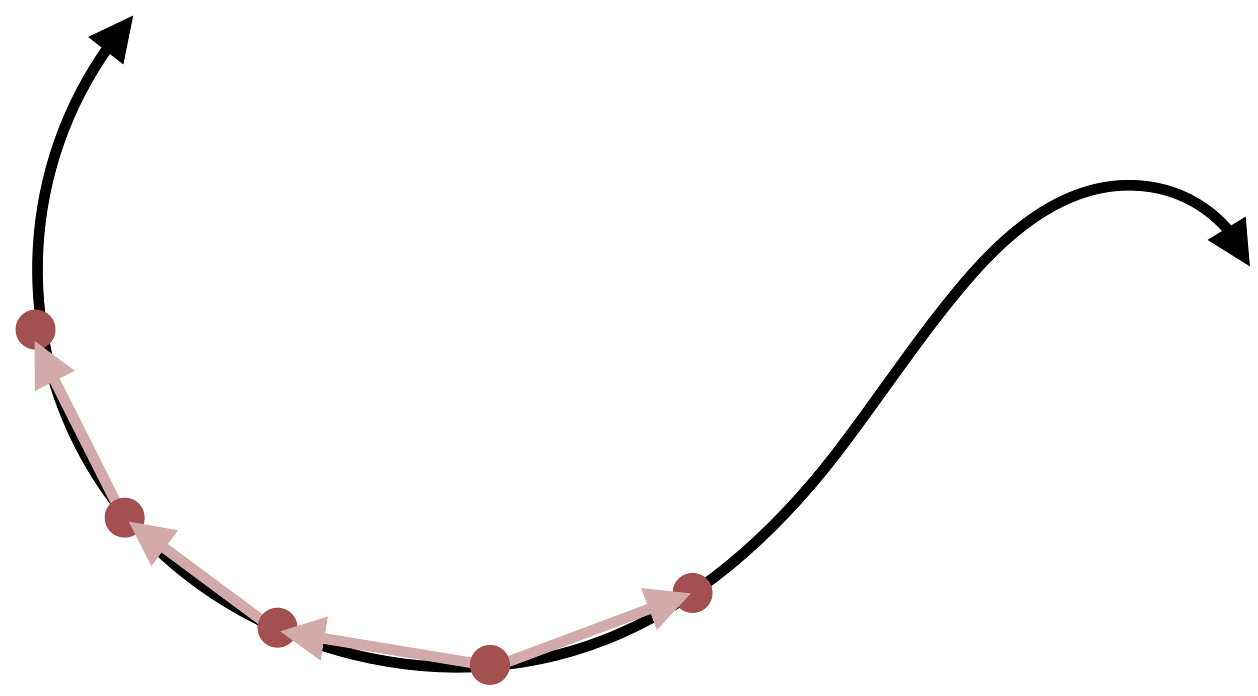
The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.



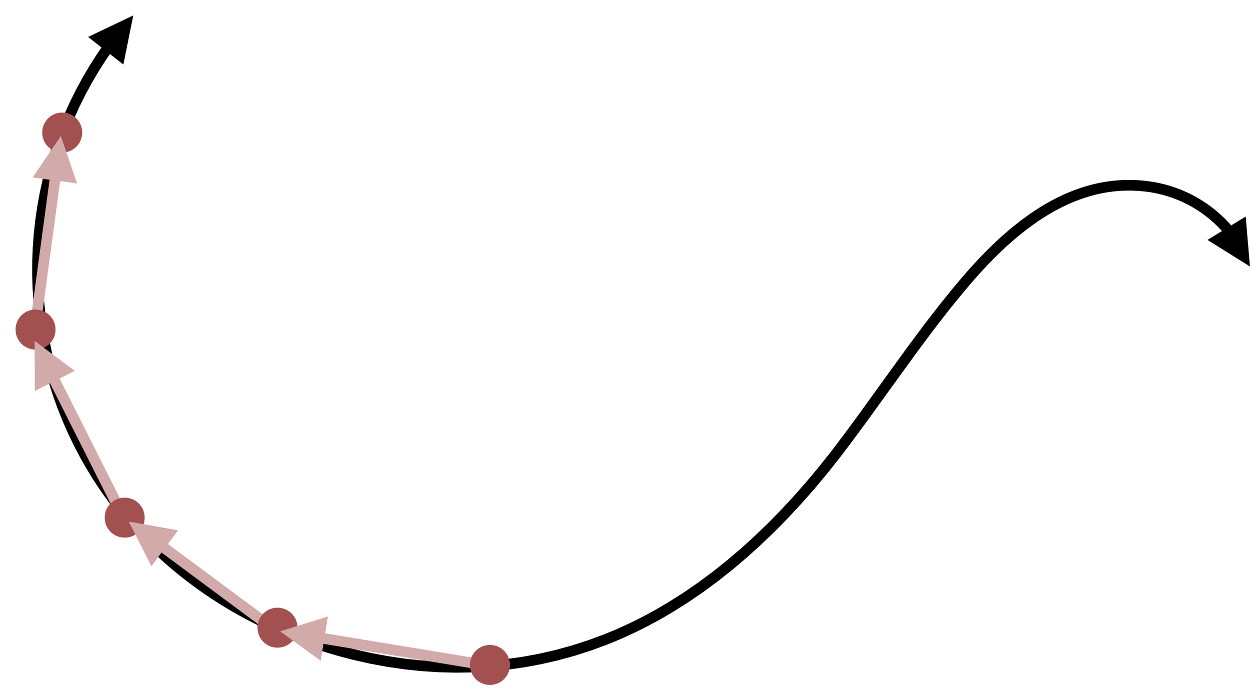
The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.



The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.



The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

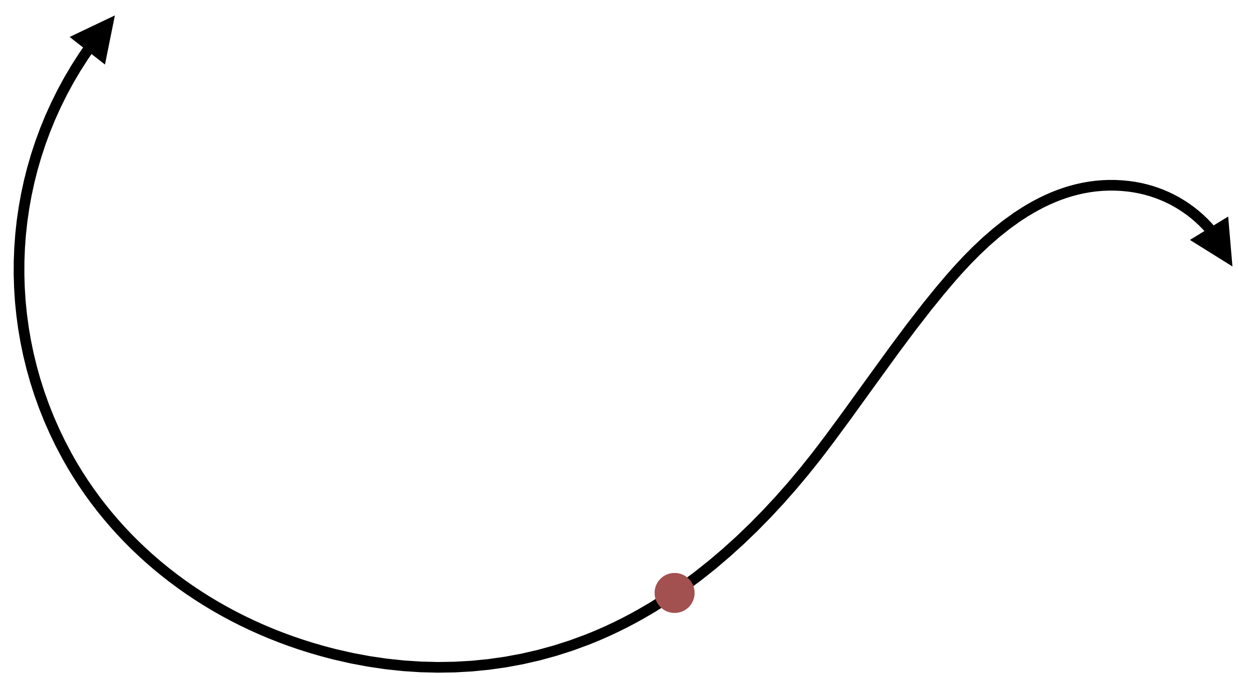
$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t}, \; \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.



$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

$$p \sim \pi(p \mid q)$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

The design of valid correction schemes is facilitated by considering entire trajectories instead of single points.

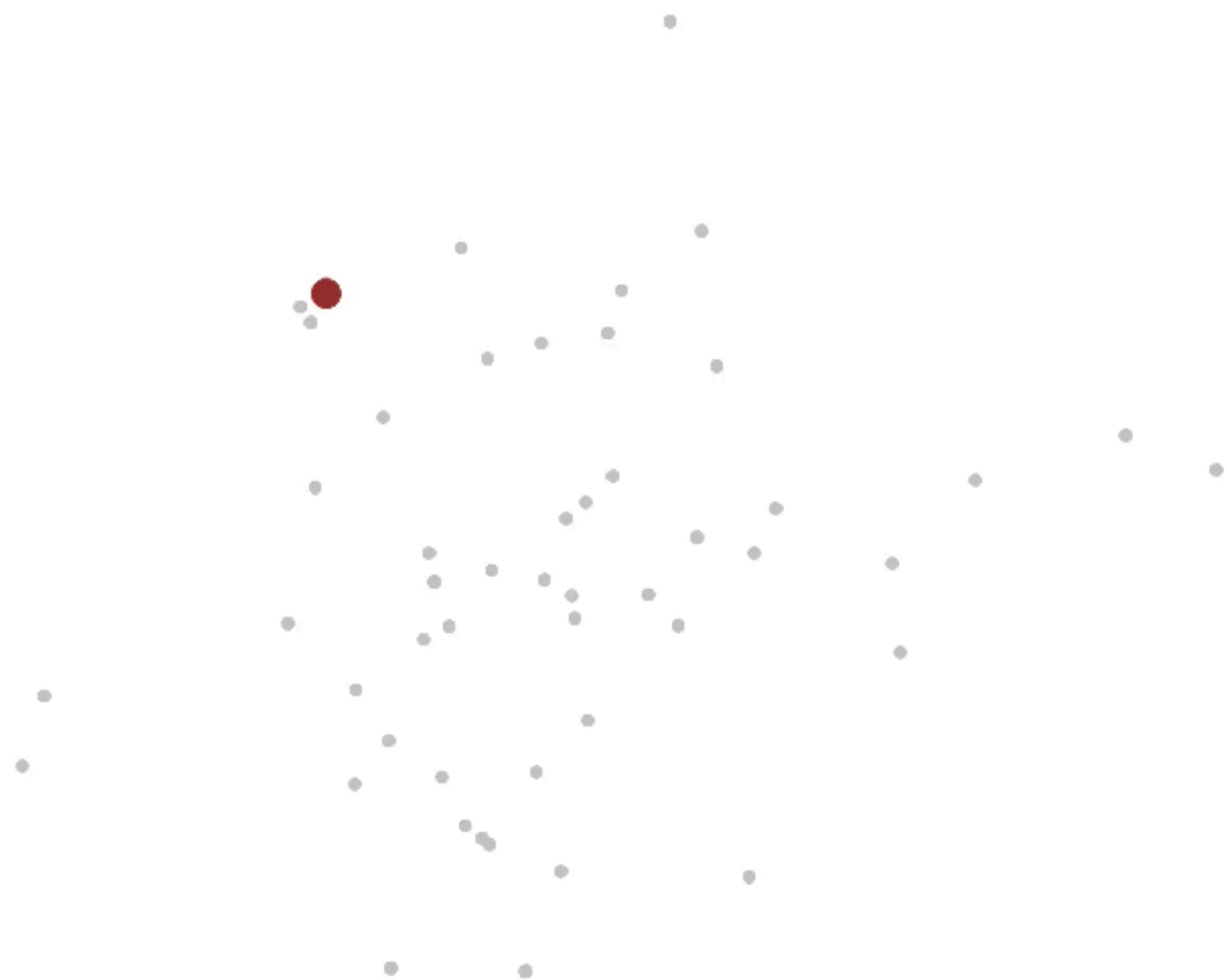
$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

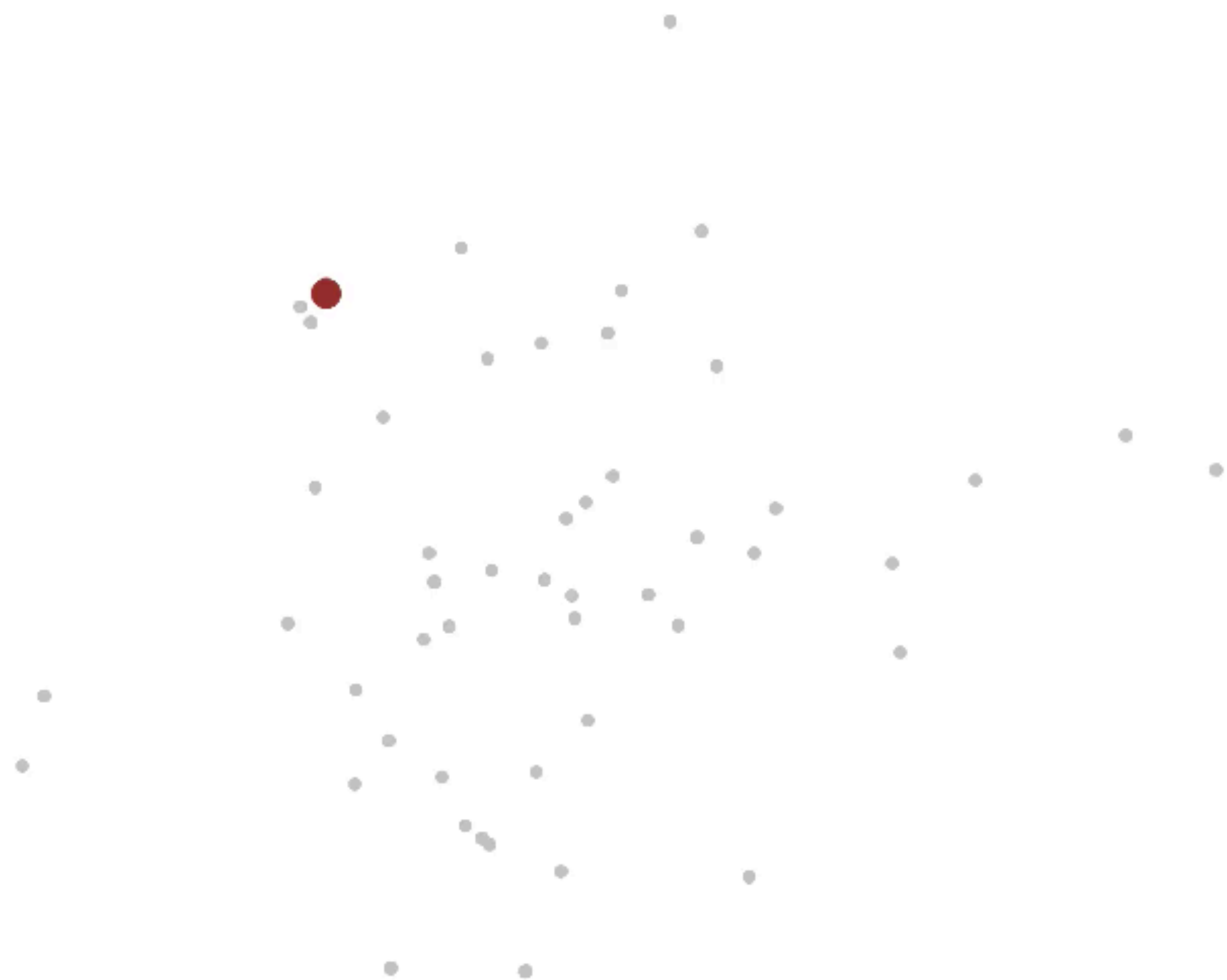
$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q',p')}$$

$$(q', p') \to q'$$

Hamiltonian Monte Carlo yields fast, and *robust*, exploration of the distributions common in practice.



Hamiltonian Monte Carlo yields fast, and *robust*, exploration of the distributions common in practice.



Theory motivates optimal implementations of the many degrees of freedom in this procedure.

$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

$$(q', p') \to q'$$

Theory motivates optimal implementations of the many degrees of freedom in this procedure.

$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

$$(q', p') \to q'$$

Theory motivates optimal implementations of the many degrees of freedom in this procedure.

$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,\boldsymbol{T}/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

$$(q', p') \to q'$$

Theory motivates optimal implementations of the many degrees of freedom in this procedure.

$$p \sim \pi(p \mid q)$$

$$\mathfrak{t} \sim U\left(\mathfrak{T}_{q,p,T/\epsilon}\right)$$

$$(q', p') \sim \mathfrak{t},\ \mathbb{P}(q', p') \propto e^{-H(q', p')}$$

$$(q', p') \to q'$$