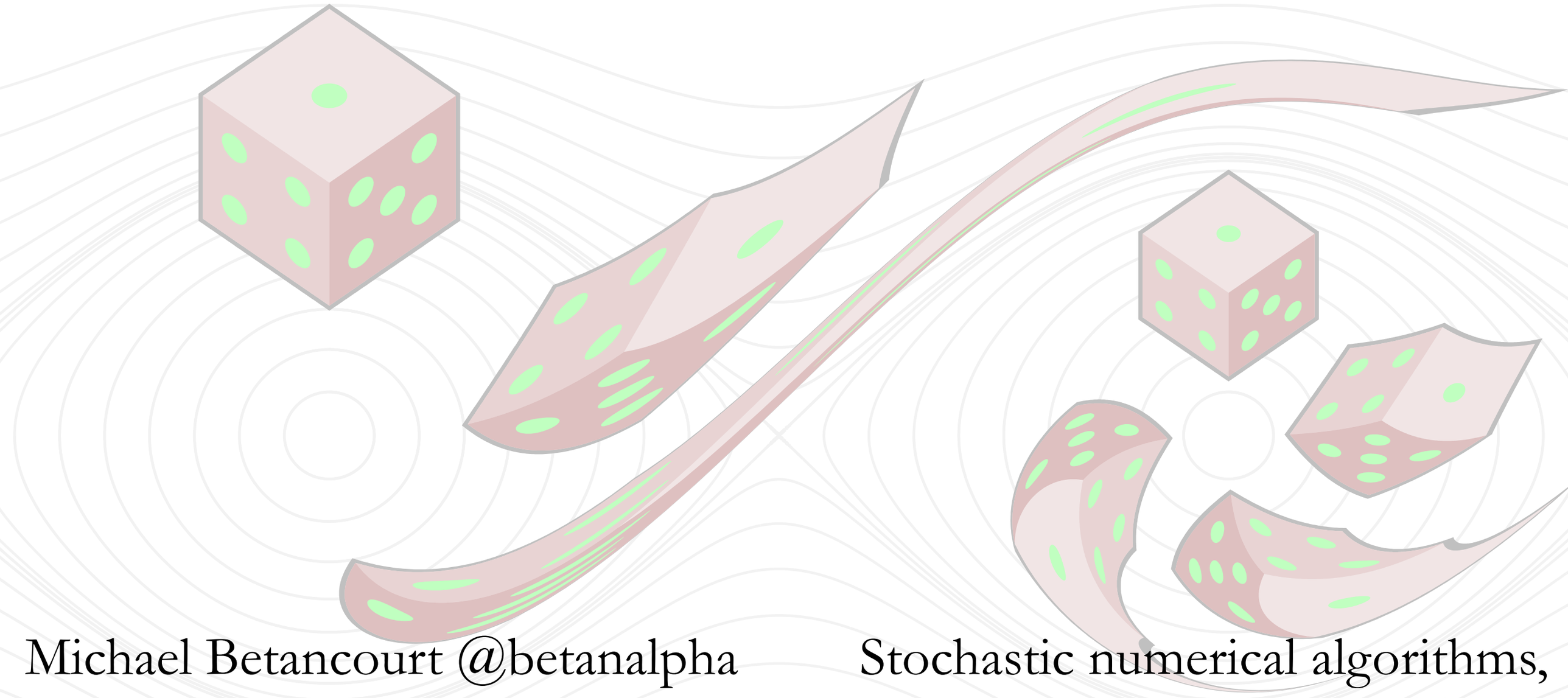
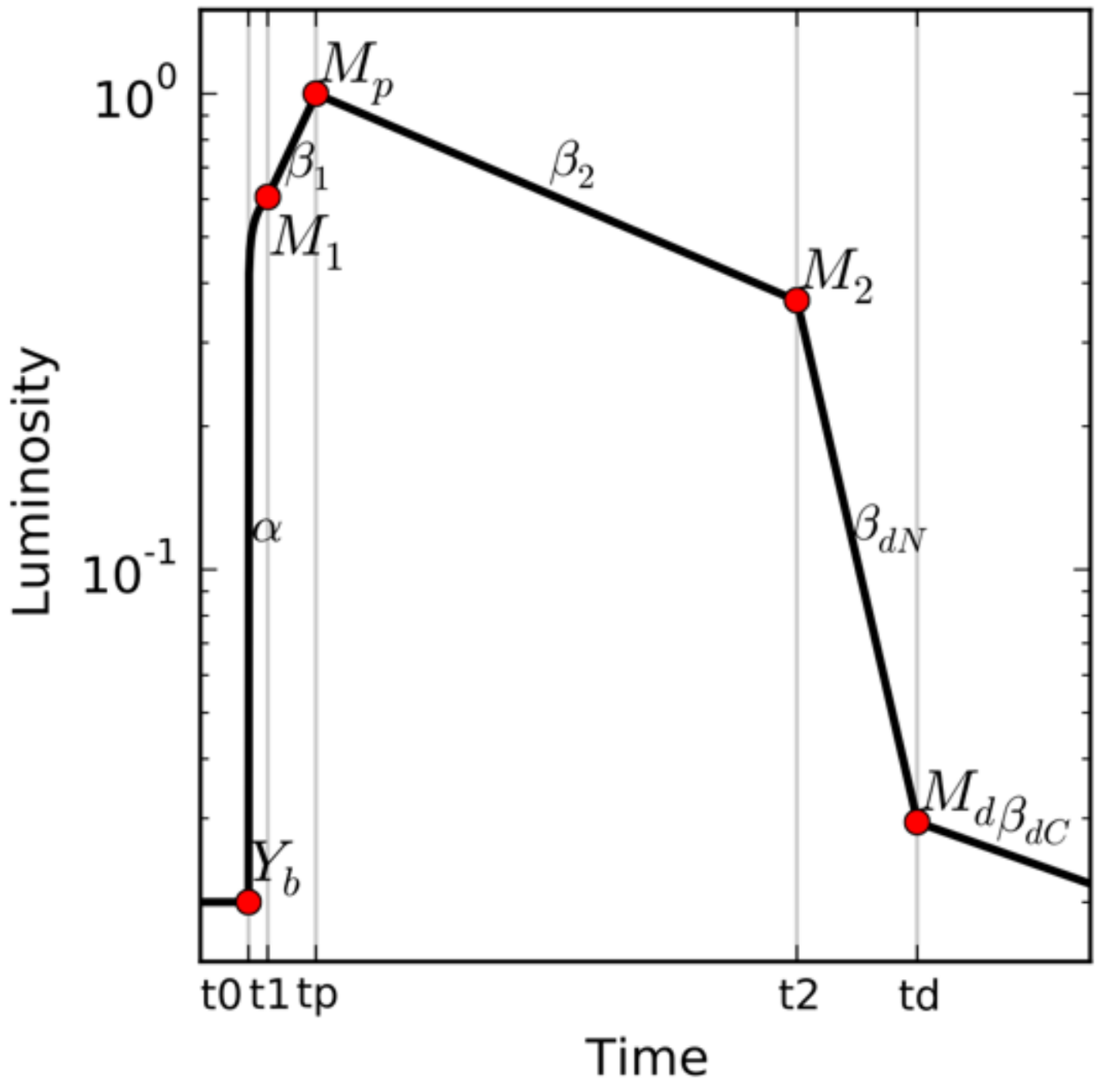
Scalable Bayesian Inference with Hamiltonian Monte Carlo



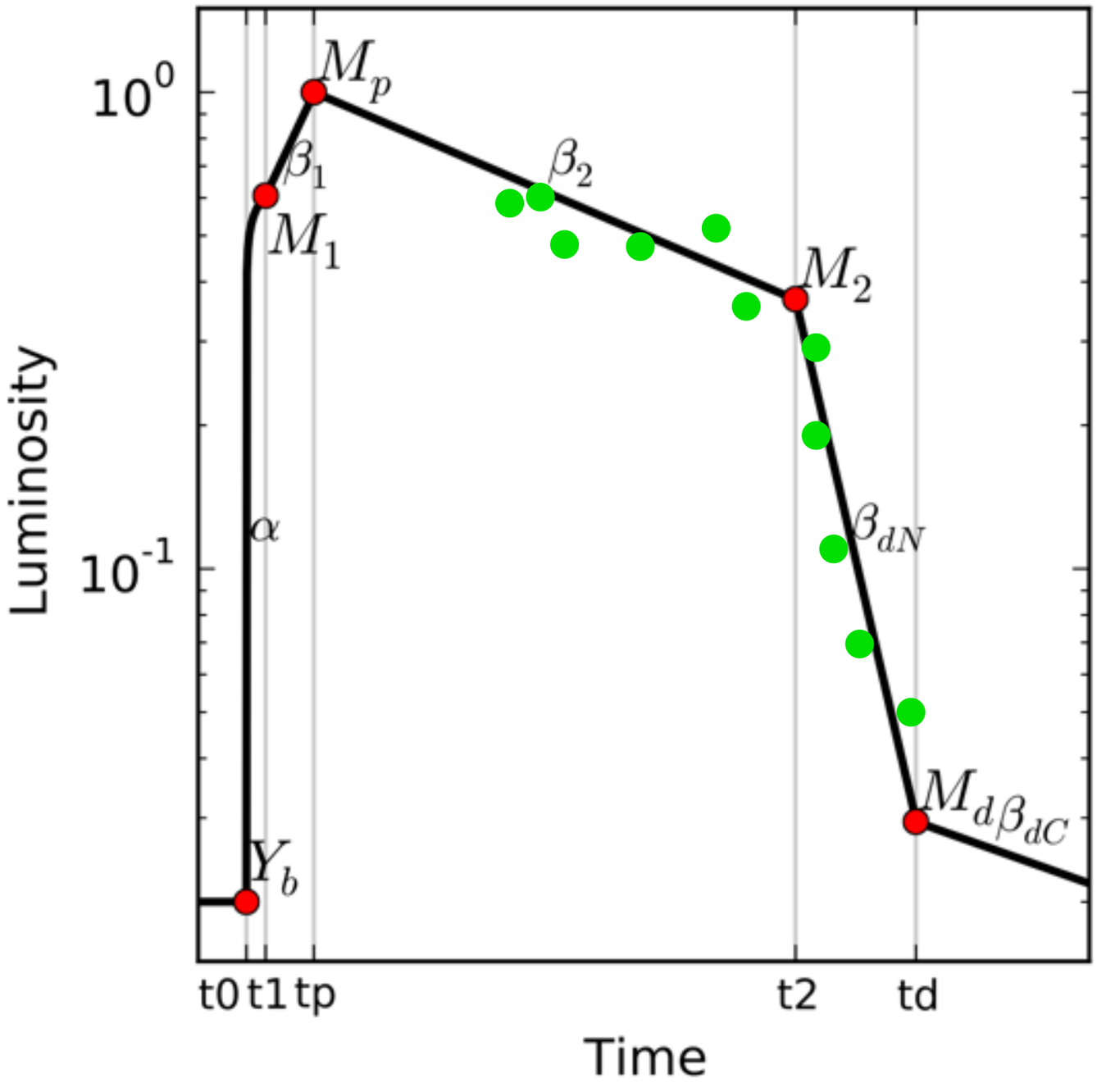
Michael Betancourt @betanalpha
Centre for Research
in Statistical Methodology,
University of Warwick

Stochastic numerical algorithms,
multiscale modeling and high-
dimensional data analytics,
ICERM, Providence, July 21, 2016

Big data is a messy business...

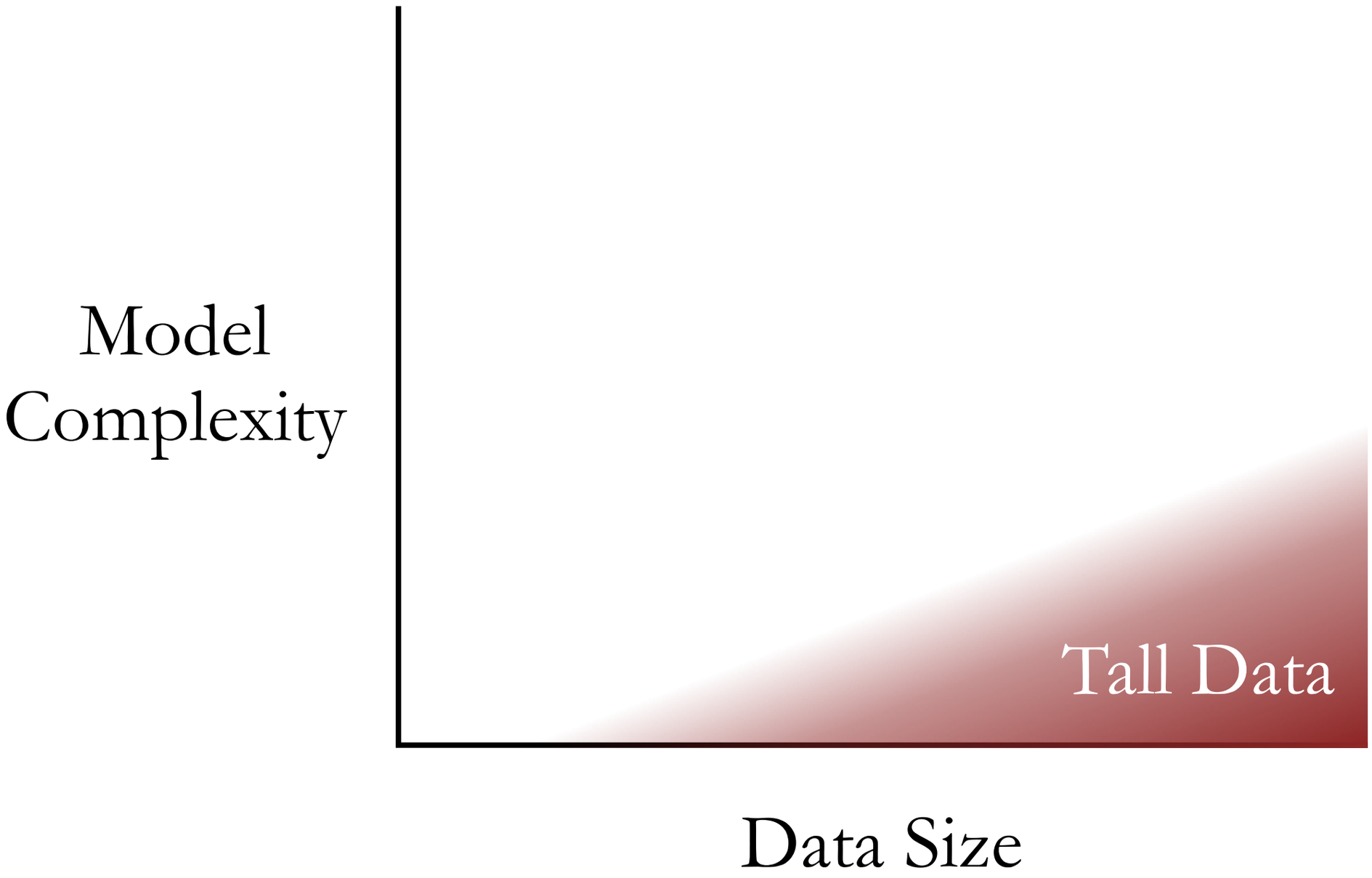


Big data is a messy business...

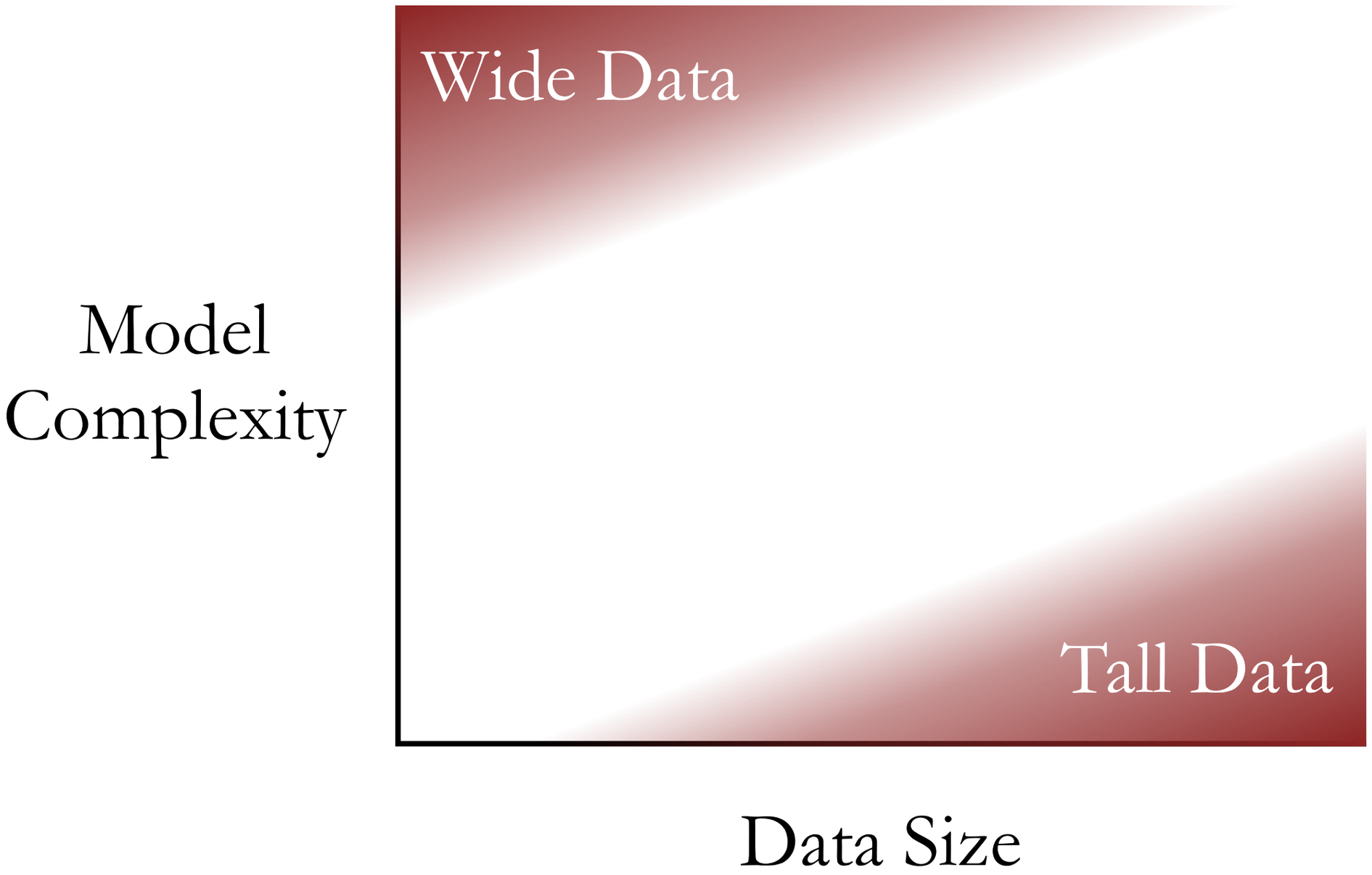


In order to build a complete statistical analysis we also need to consider a *model* of the structure in the data.

In order to build a complete statistical analysis we also need to consider a *model* of the structure in the data.



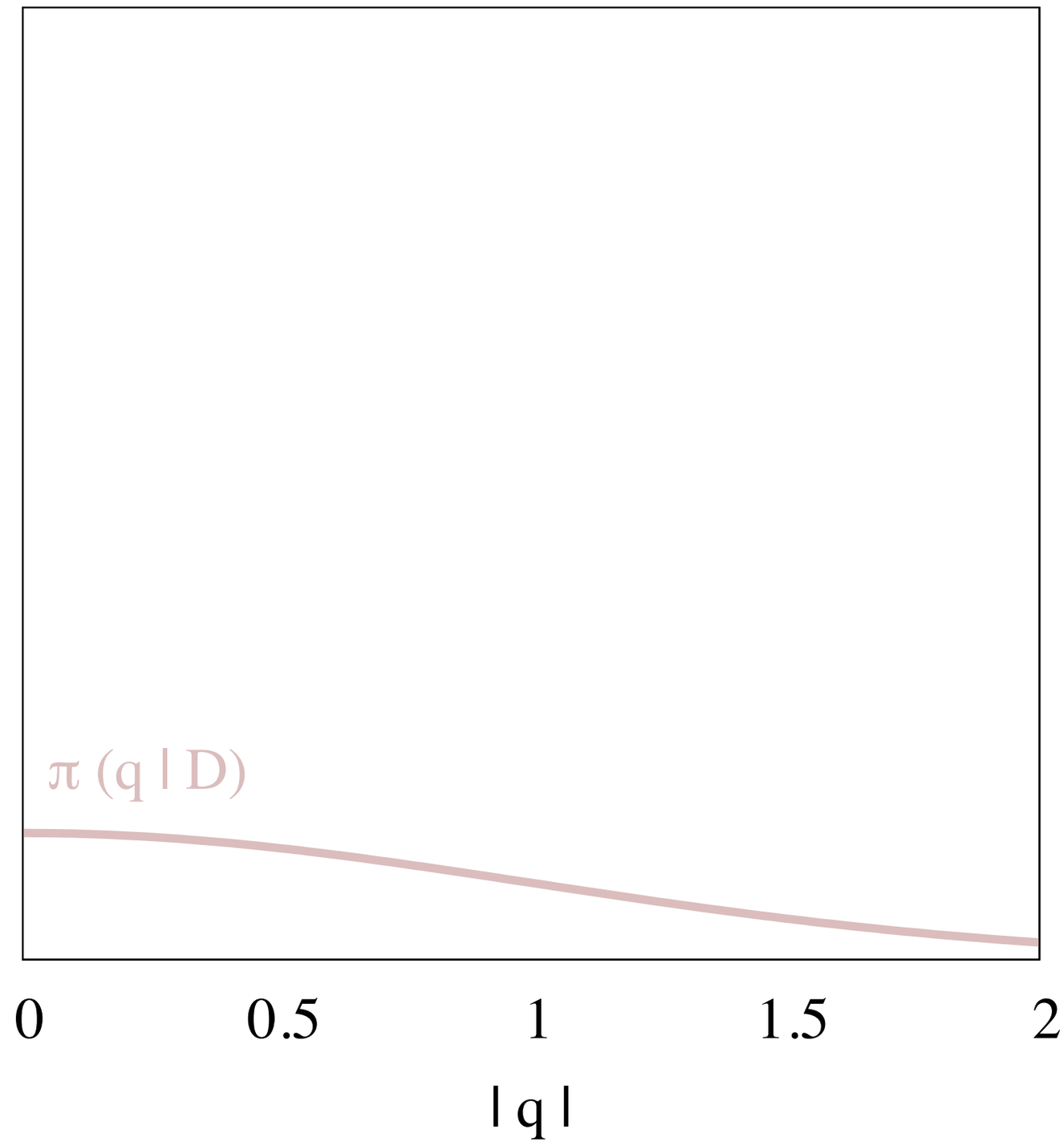
In order to build a complete statistical analysis we also need to consider a *model* of the structure in the data.



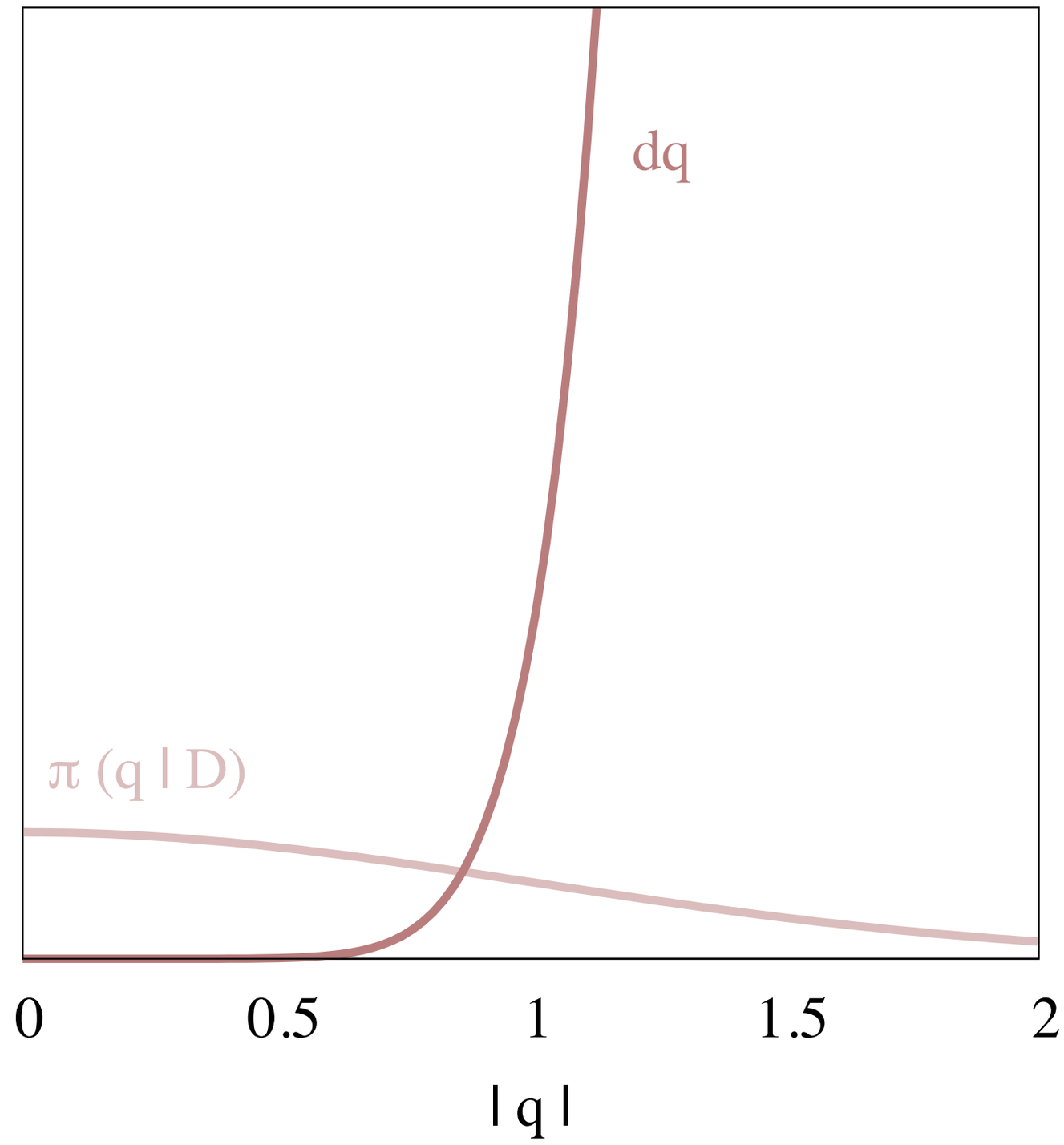
In Bayesian inference, the computational challenge reduces to estimating high-dimensional expectations.

$$\mathbb{E}_\pi[f] = \int \mathrm{d}q\, \pi(q|\mathcal{D})\, f(q)$$

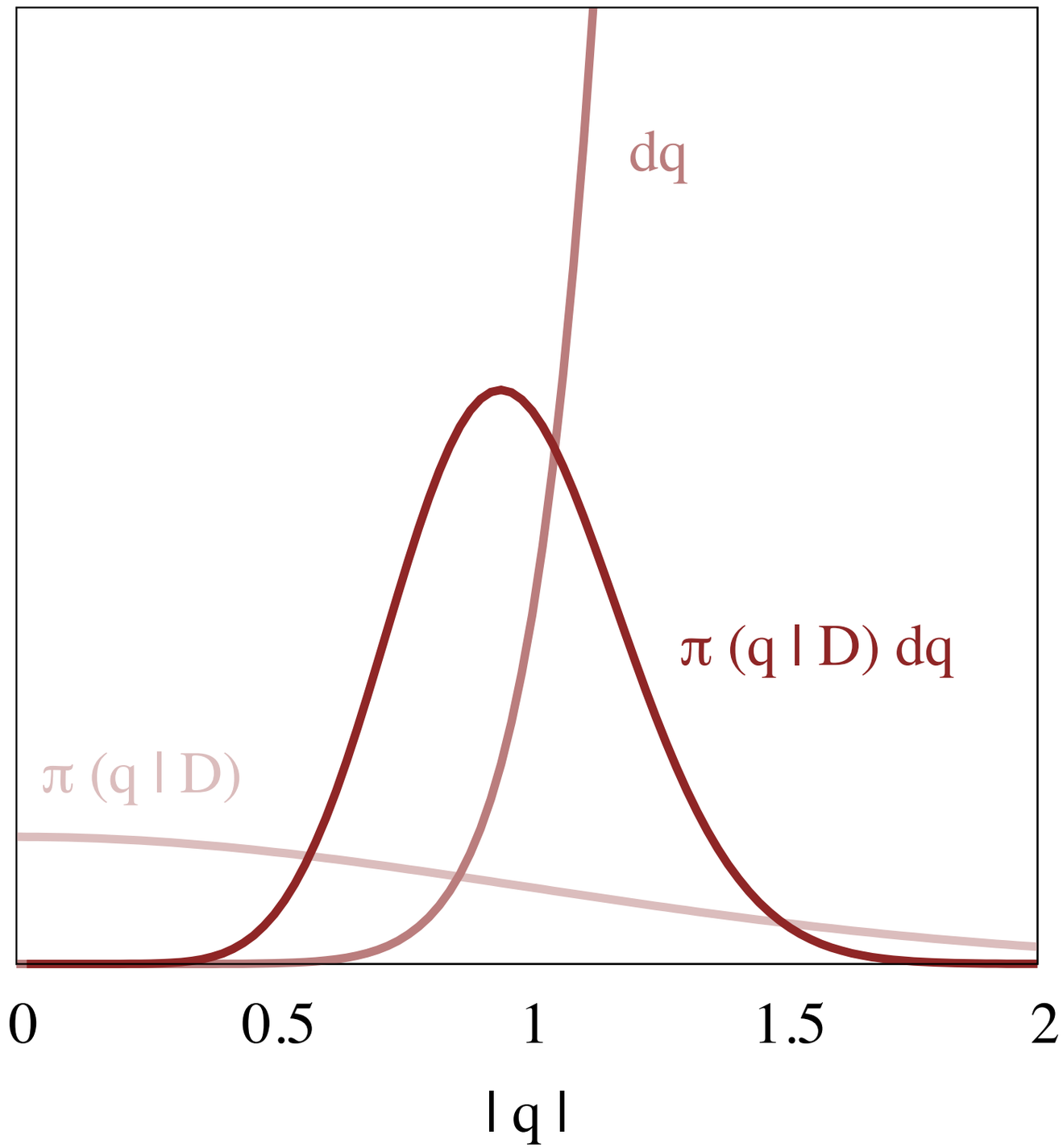
Contributions to these expectations, however, come not from probability *density* but rather probability *mass*.



Contributions to these expectations, however, come not from probability *density* but rather probability *mass*.



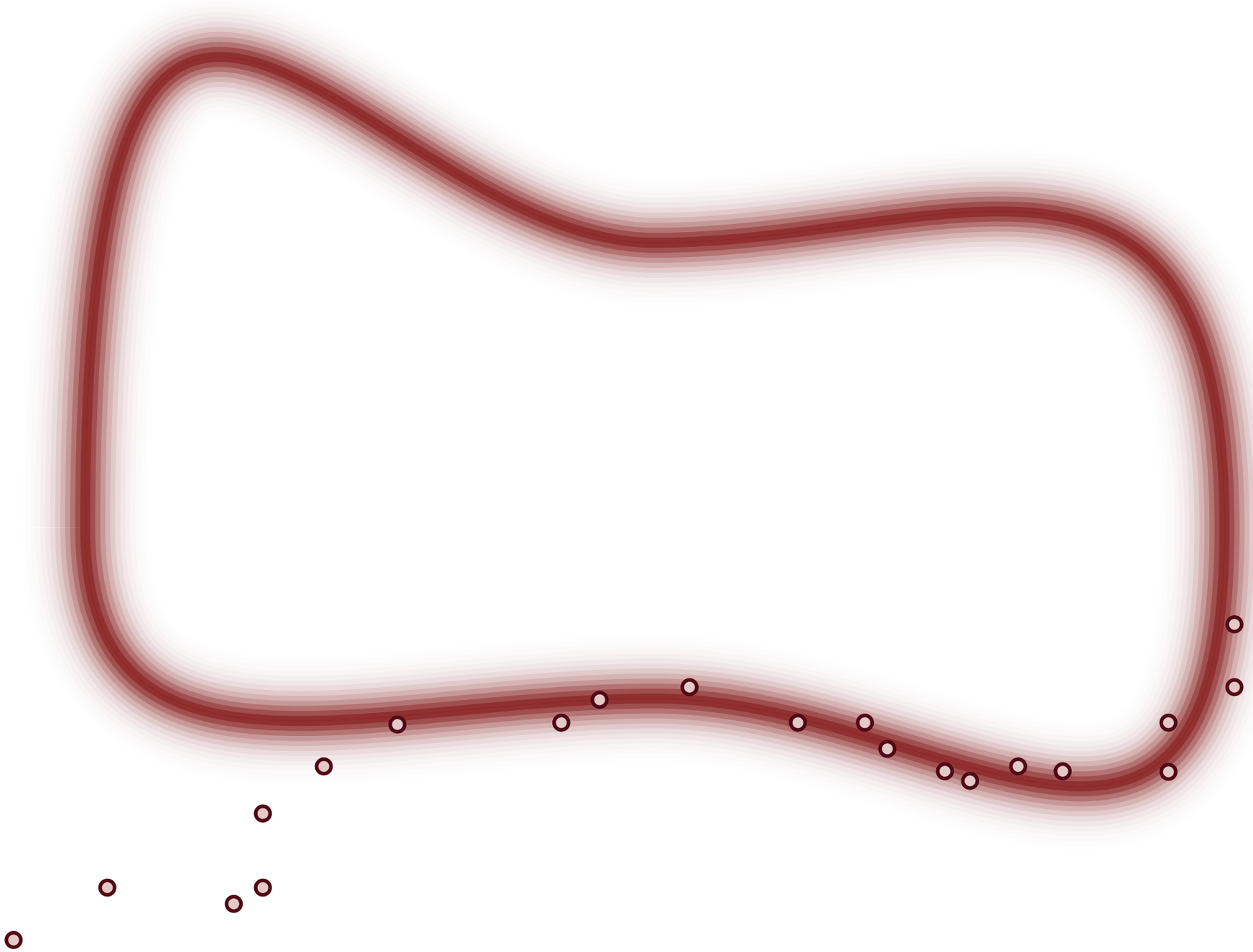
Contributions to these expectations, however, come not from probability *density* but rather probability *mass*.



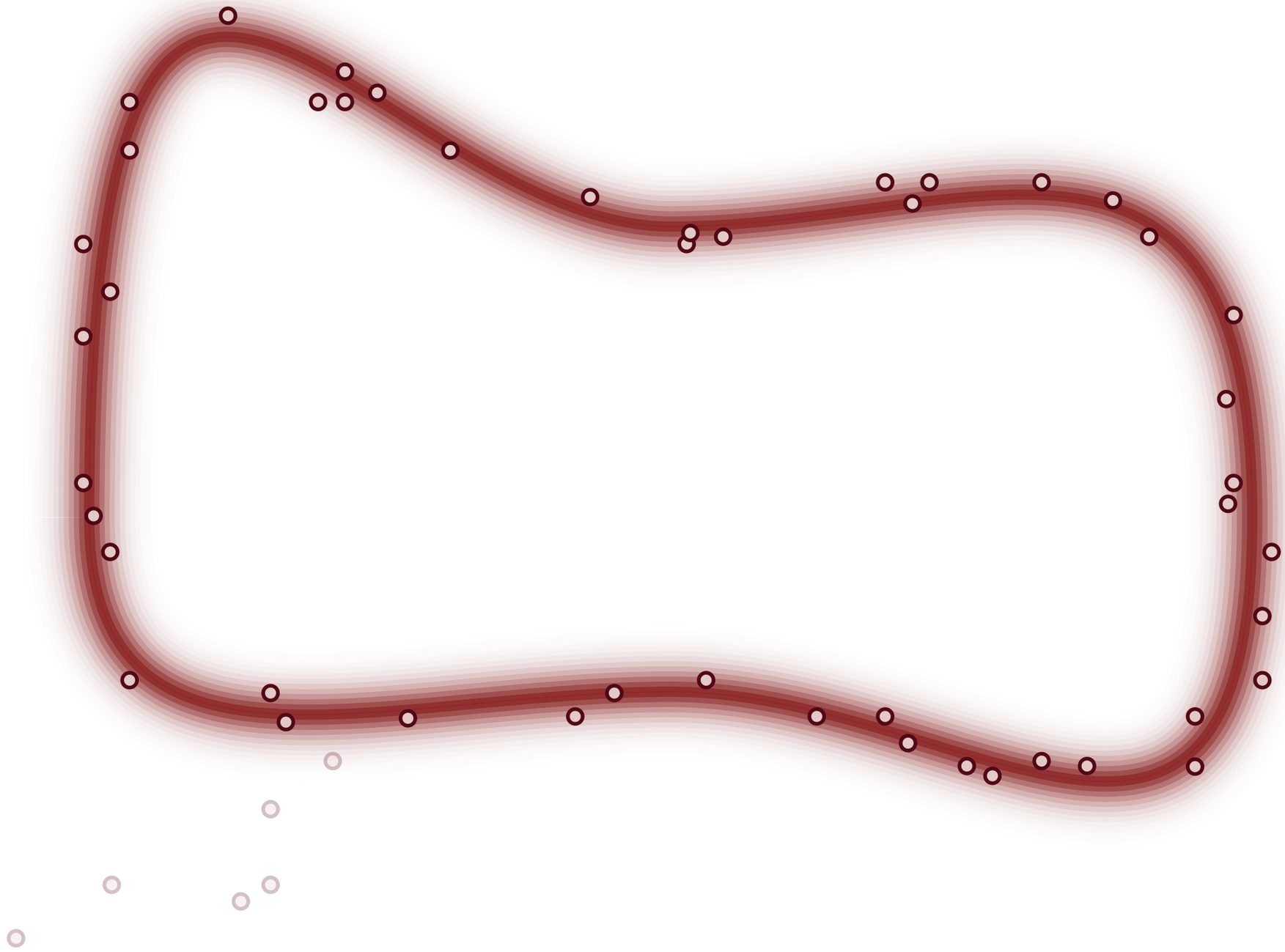
The concentration of probability mass into a narrow typical set frustrates the accurate estimation of integrals.

Markov chains provide a generic and practical scheme for finding and then exploring this typical set.

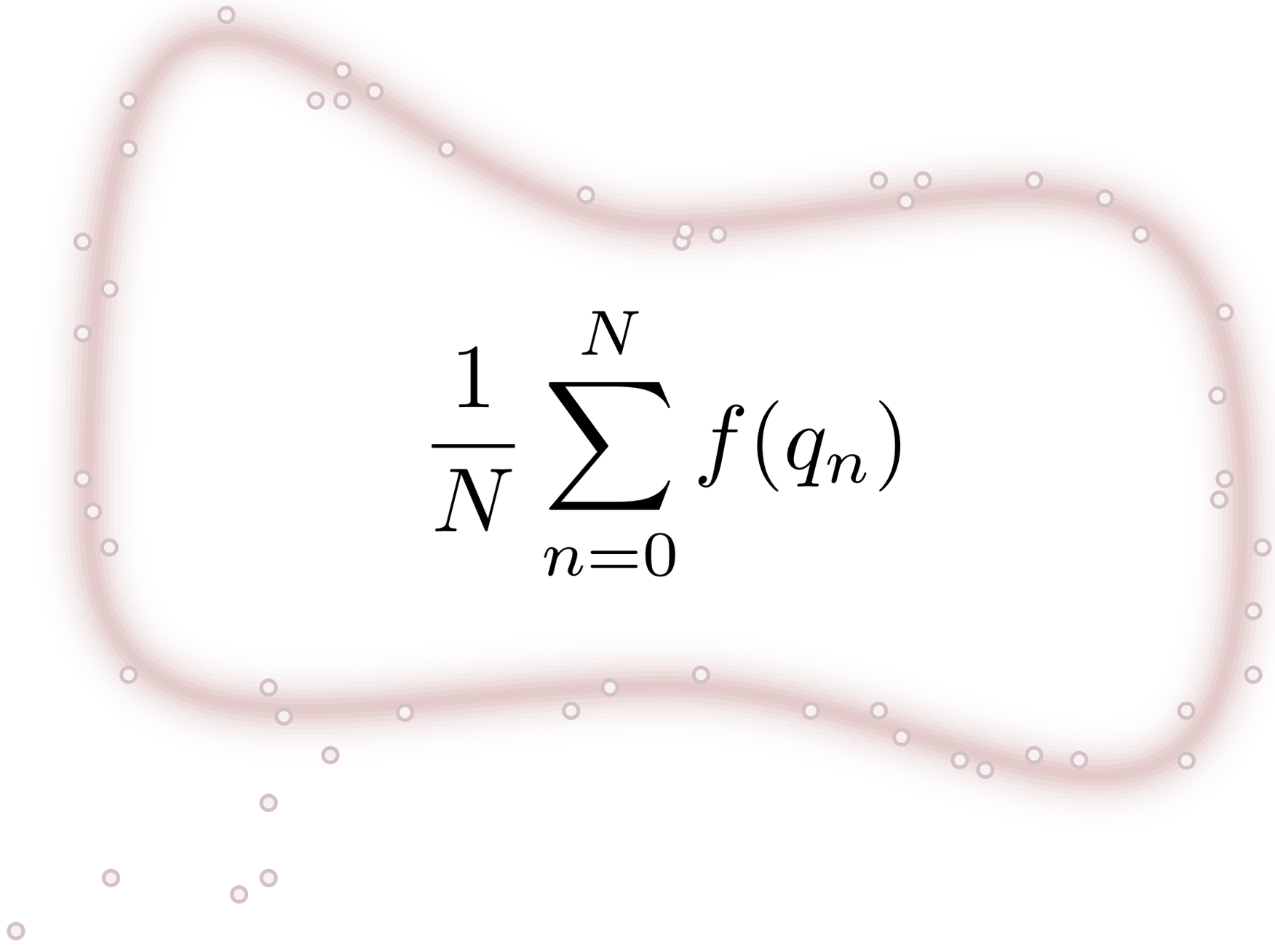
Markov chains provide a generic and practical scheme for finding and then exploring this typical set.



If run long enough the Markov chain defines consistent *Markov Chain Monte Carlo estimators.*

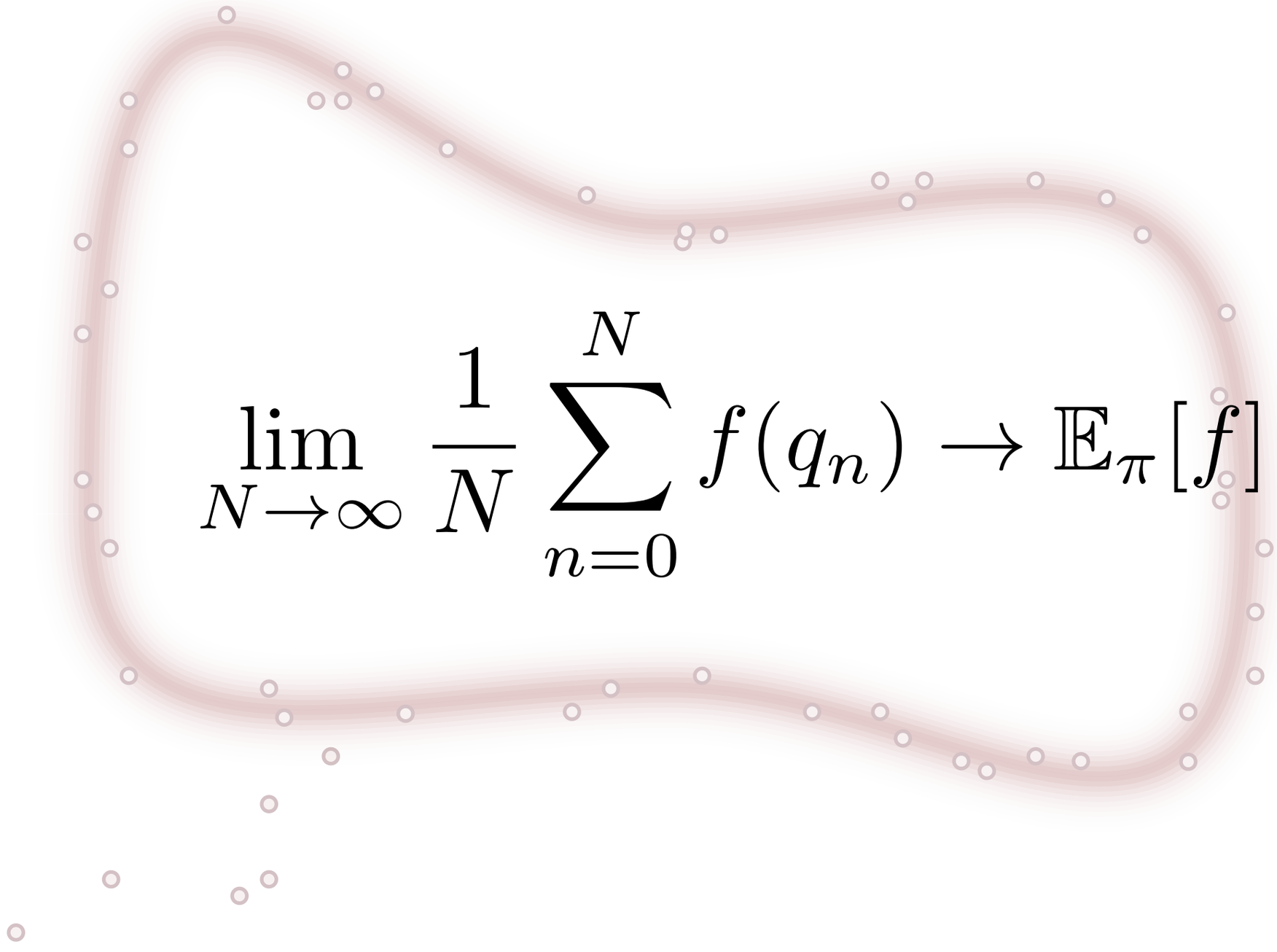


If run long enough the Markov chain defines consistent *Markov Chain Monte Carlo estimators*.



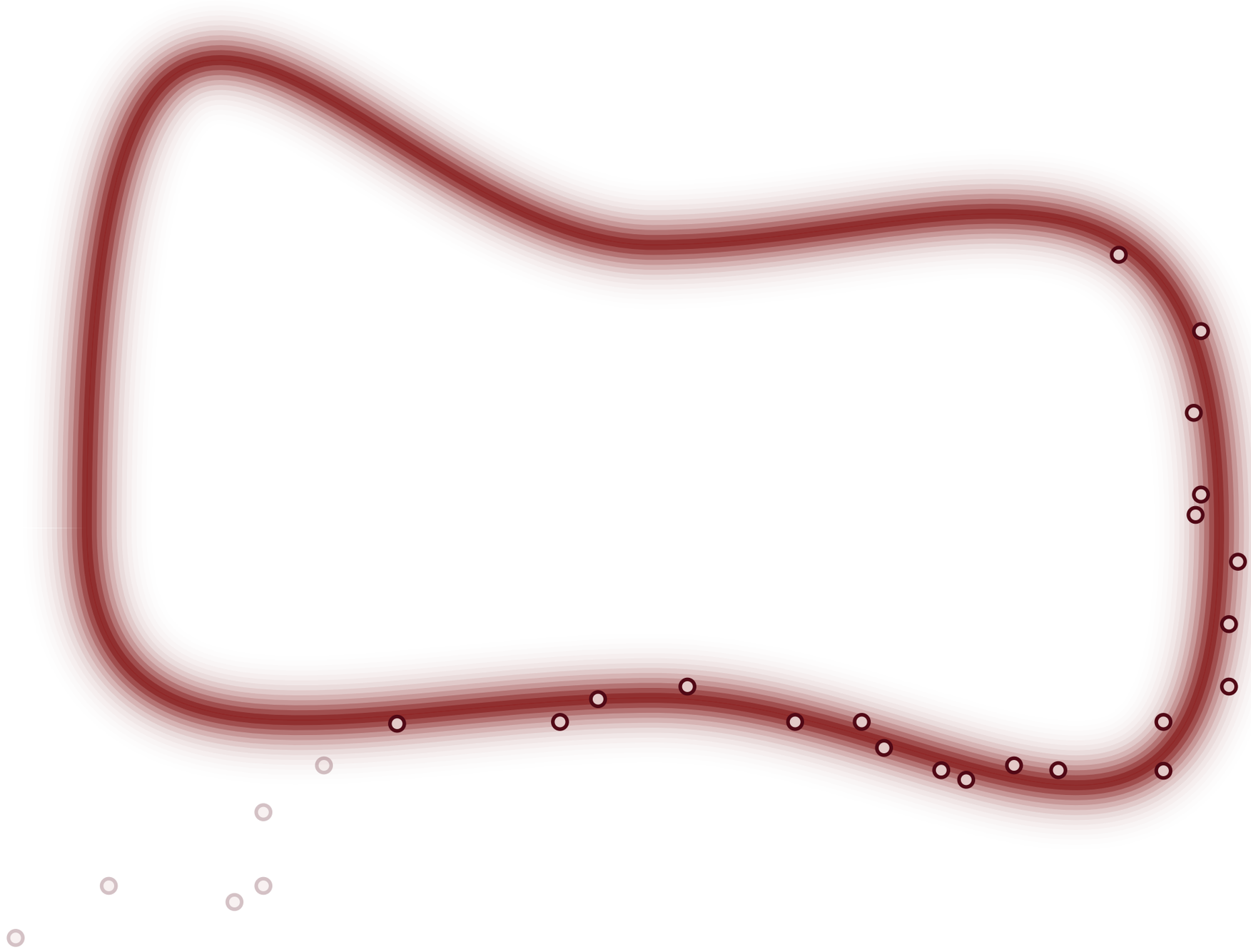
$$\frac{1}{N}\sum_{n=0}^{N} f(q_n)$$

If run long enough the Markov chain defines consistent *Markov Chain Monte Carlo estimators.*

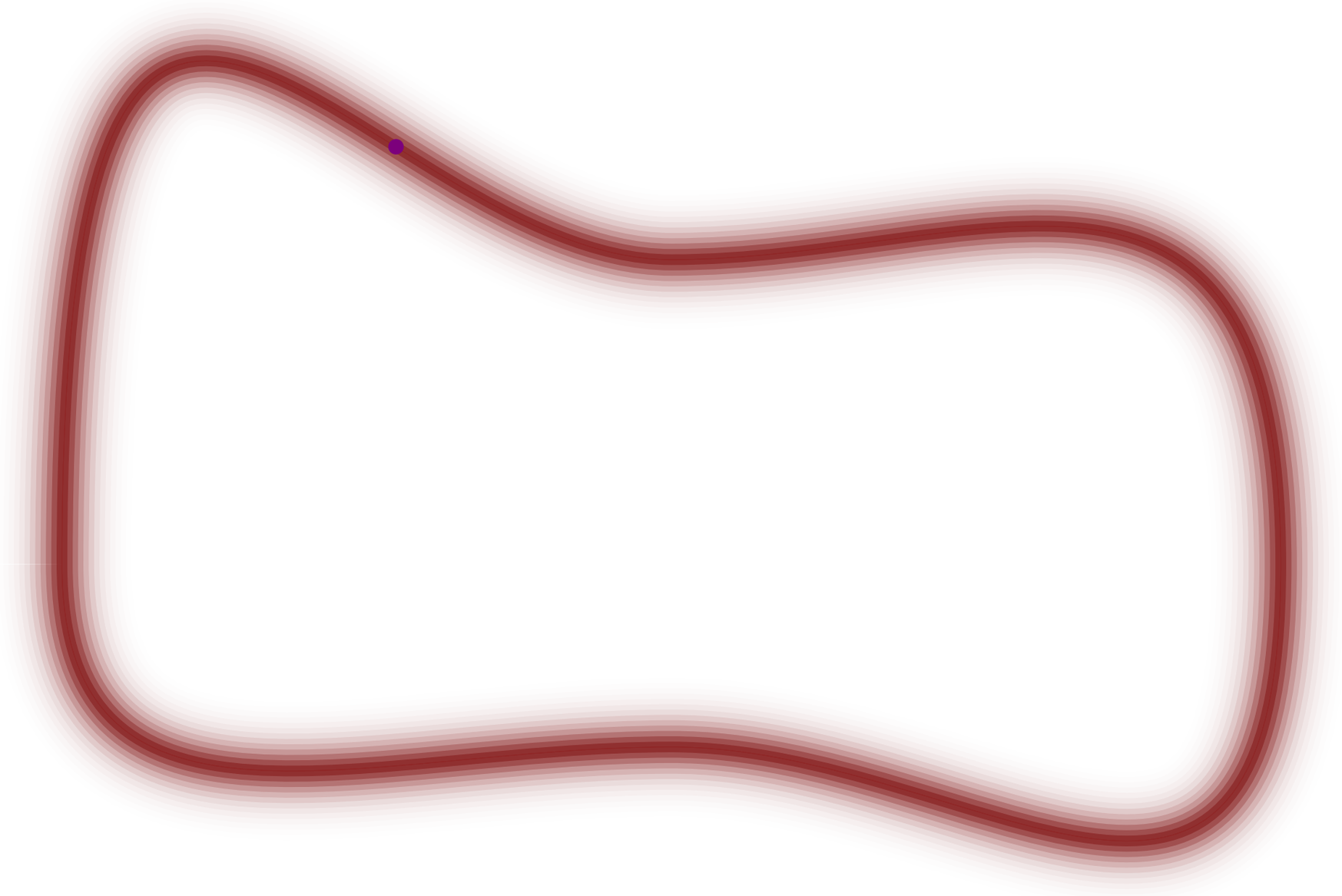


$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} f(q_n) \to \mathbb{E}_{\pi}[f]$$

But practical performance depends on how efficiently the Markov chain can explore the typical set.



In order to scale to high-dimensional target distributions we need a *coherent* exploration of the typical set.



In order to scale to high-dimensional target distributions we need a *coherent* exploration of the typical set.

Efficient exploration of the typical set is generated by measure-preserving maps on the target space.

$$\phi_t : Q \to Q$$

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

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

$$\phi_t : Q \to Q$$

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

$$\phi_t : Q \to Q$$

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

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

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

The critical ingredients to coherent exploration of the typical set are then *measure-preserving flows.*

$$\phi_t : Q \to Q$$

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

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

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

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

The critical ingredients to coherent exploration of the typical set are then *measure-preserving flows*.

$$\phi_t : Q \to Q$$

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

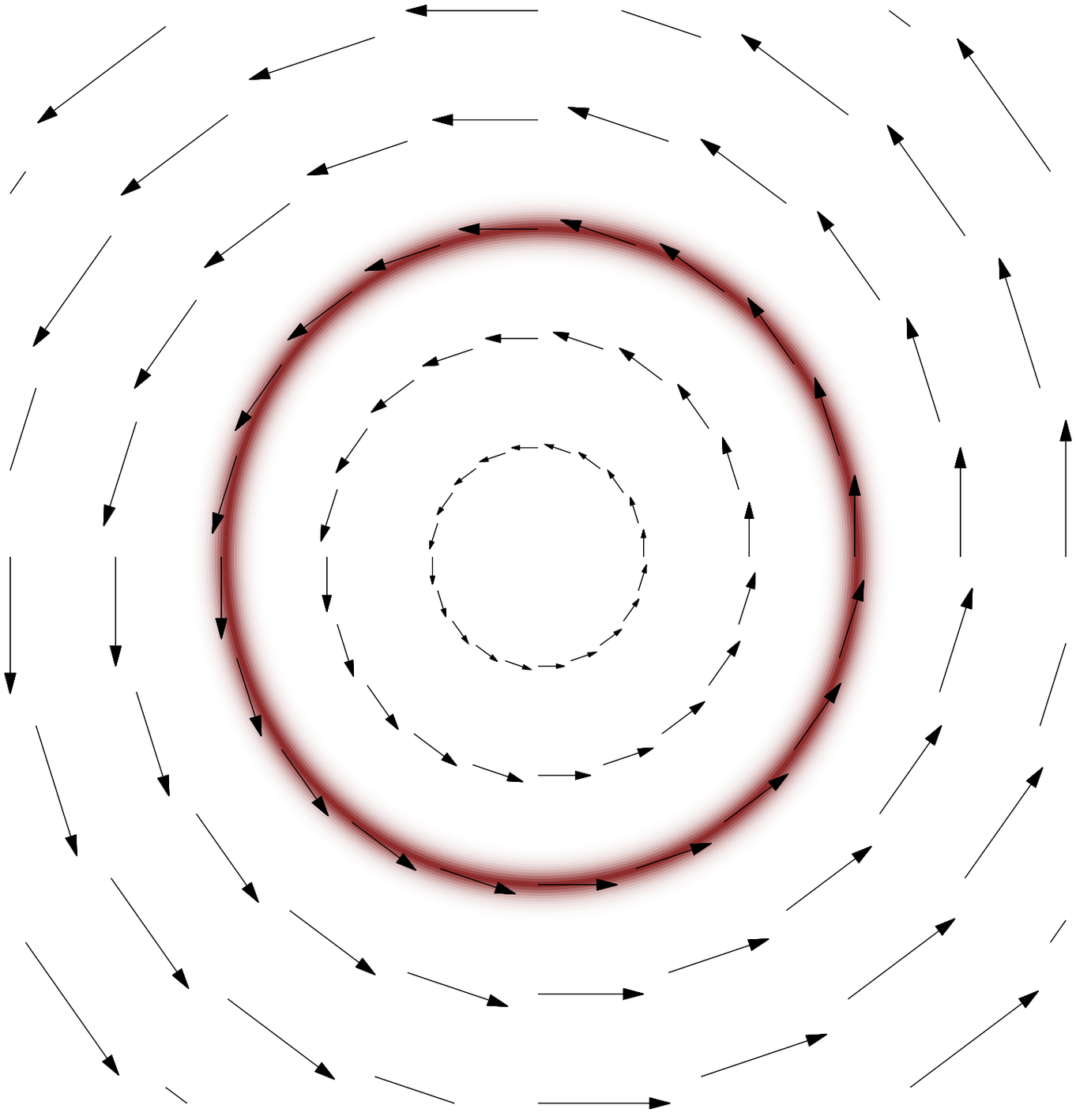
$$\phi_t^{-1} = \phi_{-t}$$

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

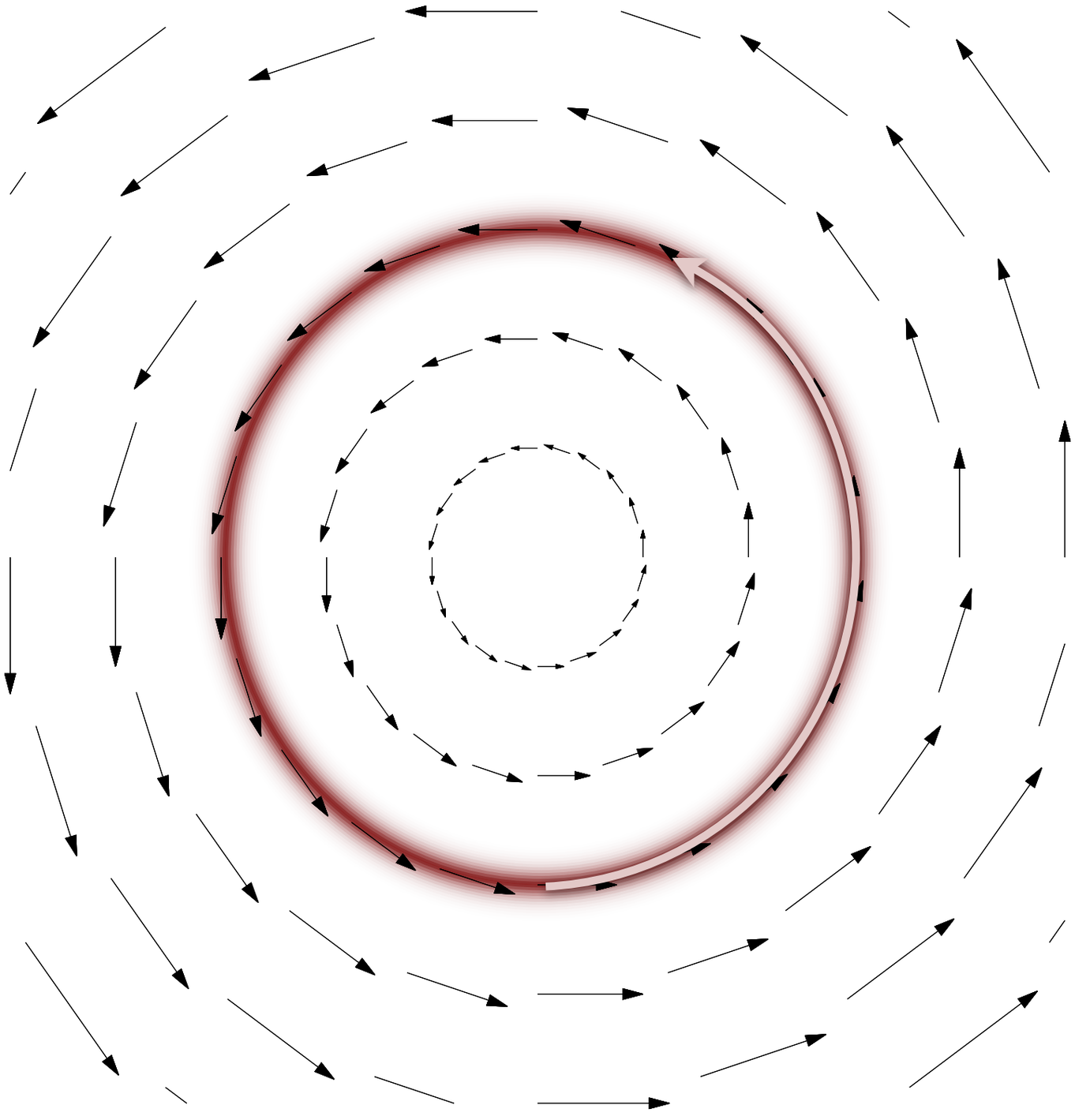
$$\left(\pi \circ \phi_t^{-1}\right)(A) = \pi(A)$$

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

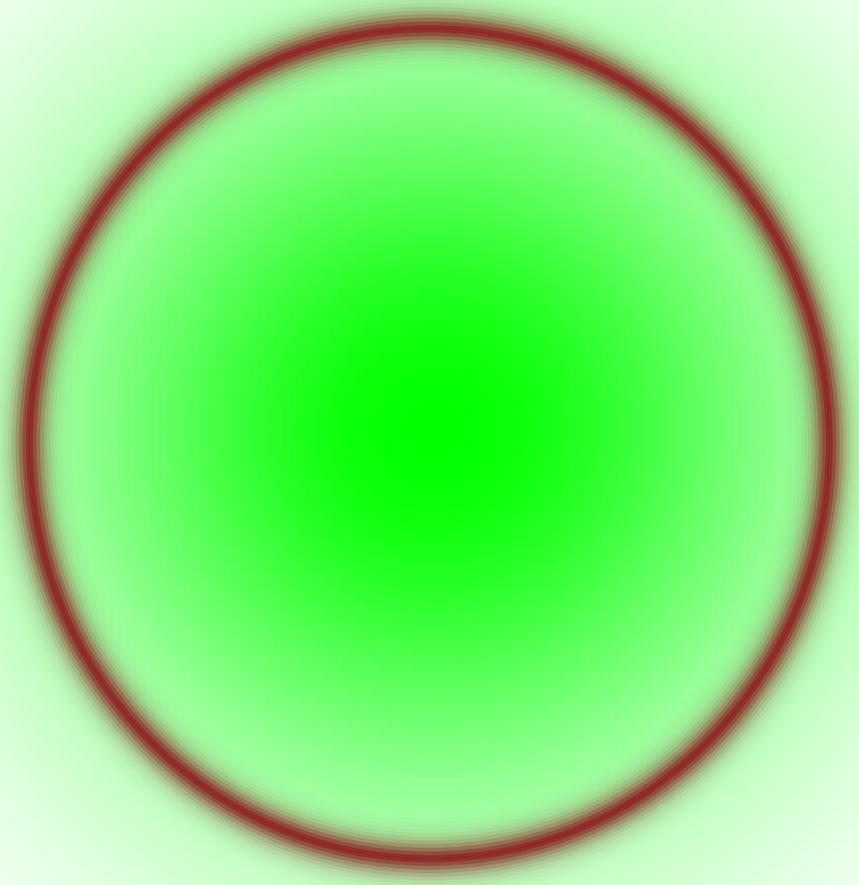
One way to construct the desired flow is to integrate along a *vector field* aligned with the typical set.



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

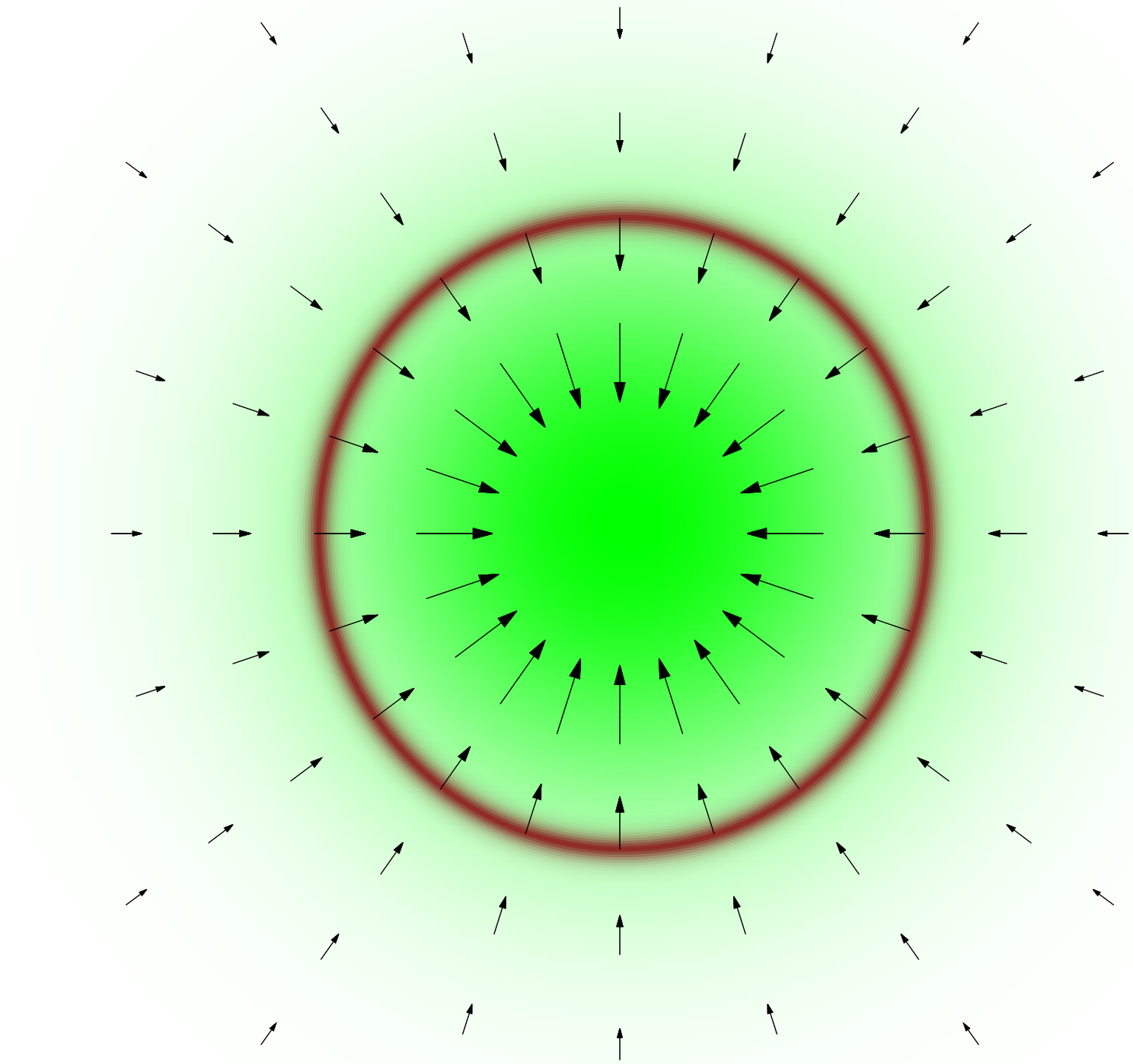


Creating the desired vector field requires transforming available vector fields, such as the gradient.



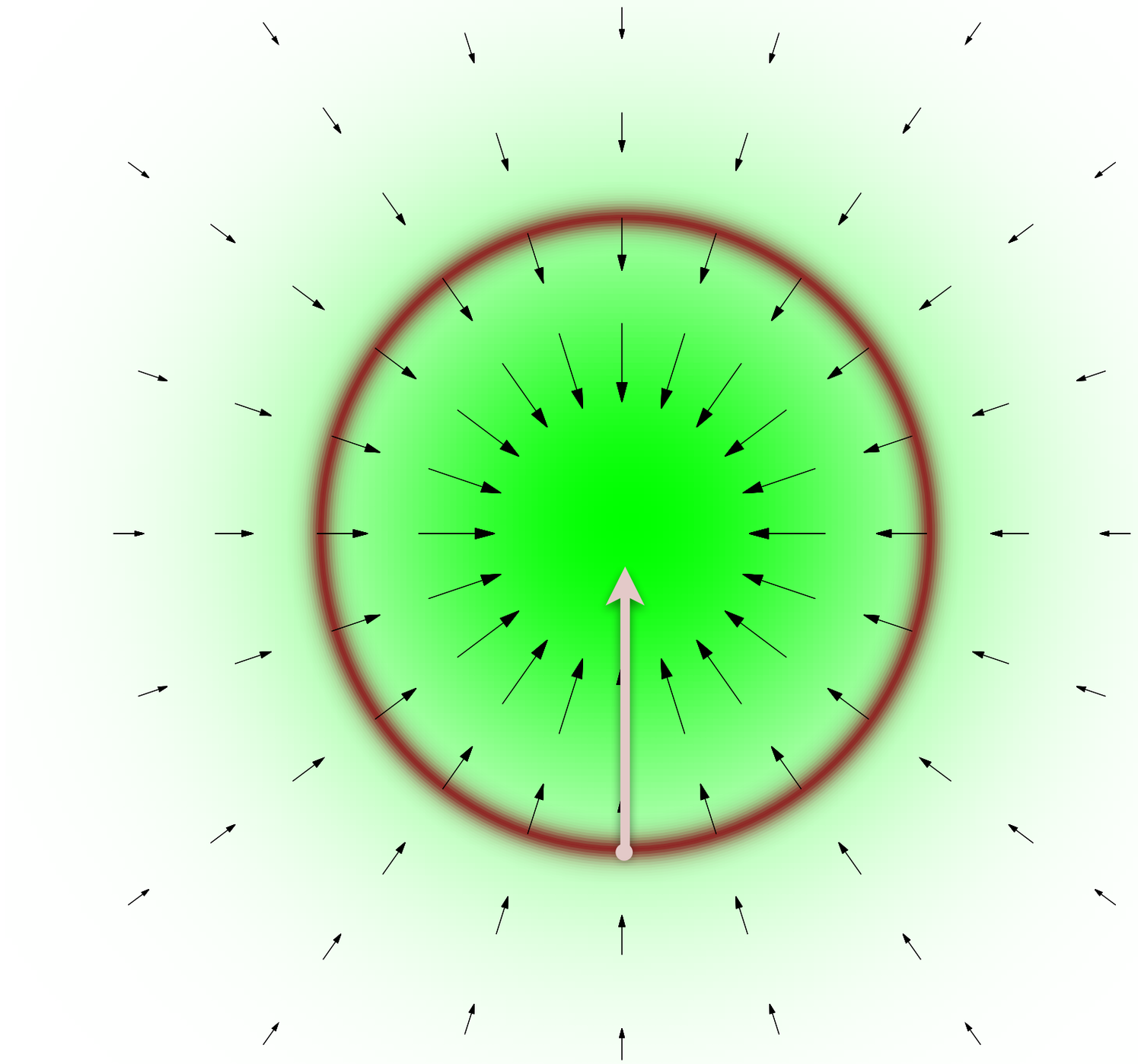
$\pi(q|\mathcal{D})$

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



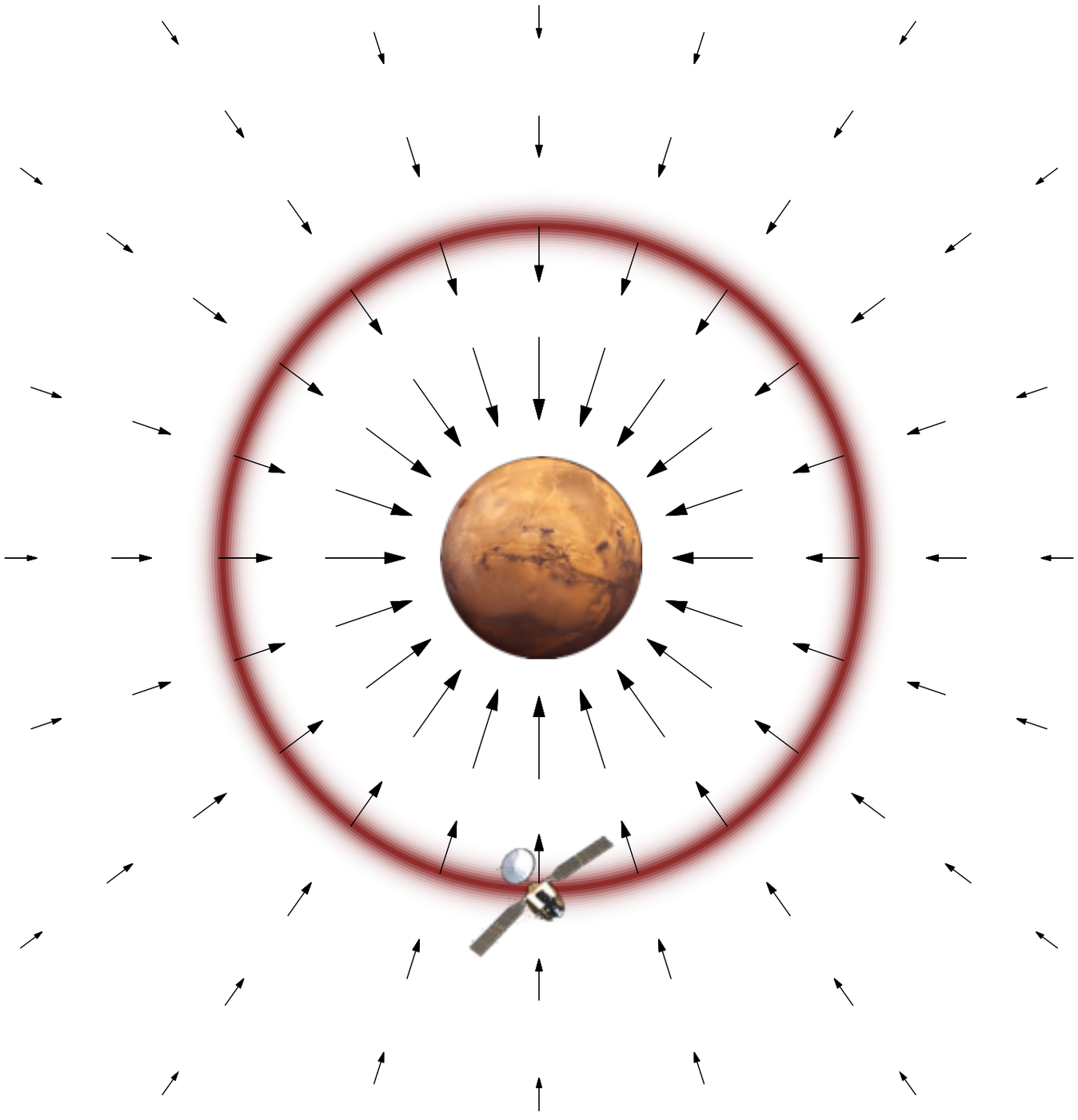
$$\frac{\partial \pi(q|\mathcal{D})}{\partial q}$$

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

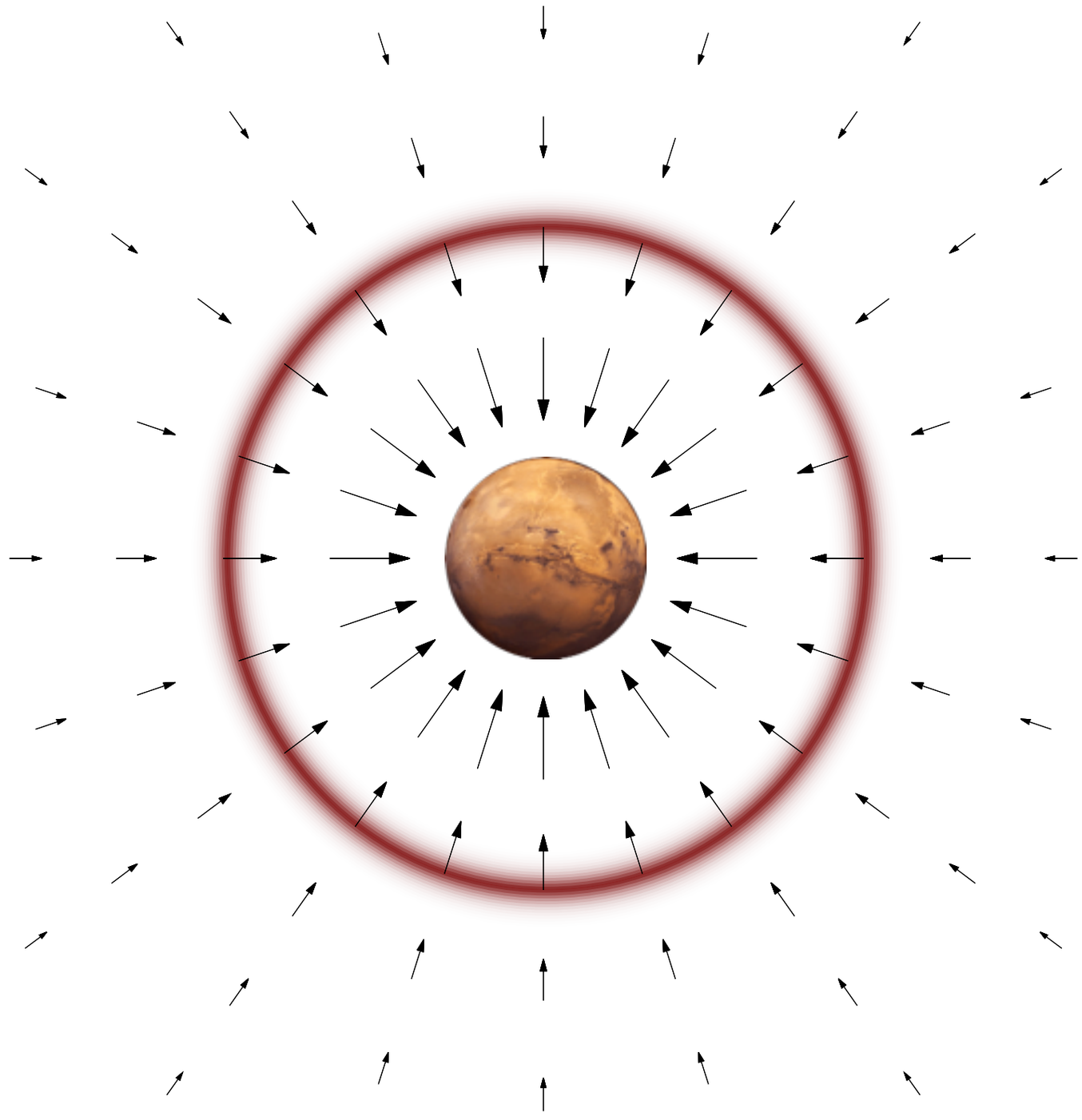


$$\frac{\partial \pi(q|\mathcal{D})}{\partial q}$$

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.

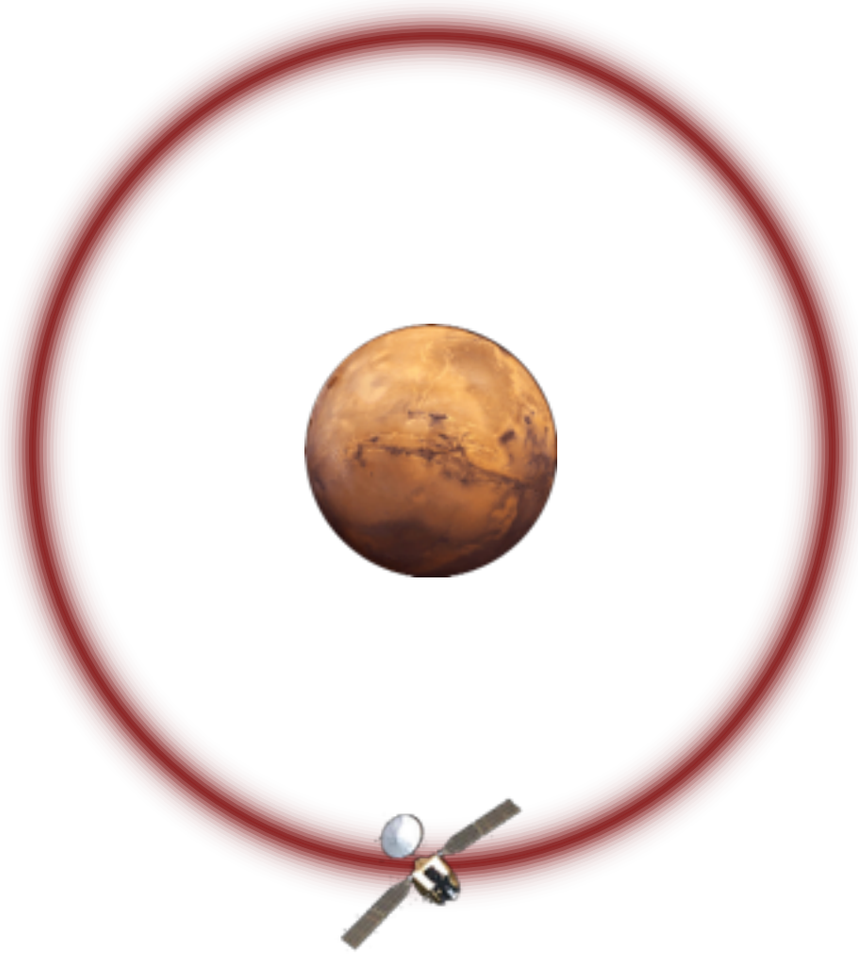



MOMENTUM
FITNESS

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.

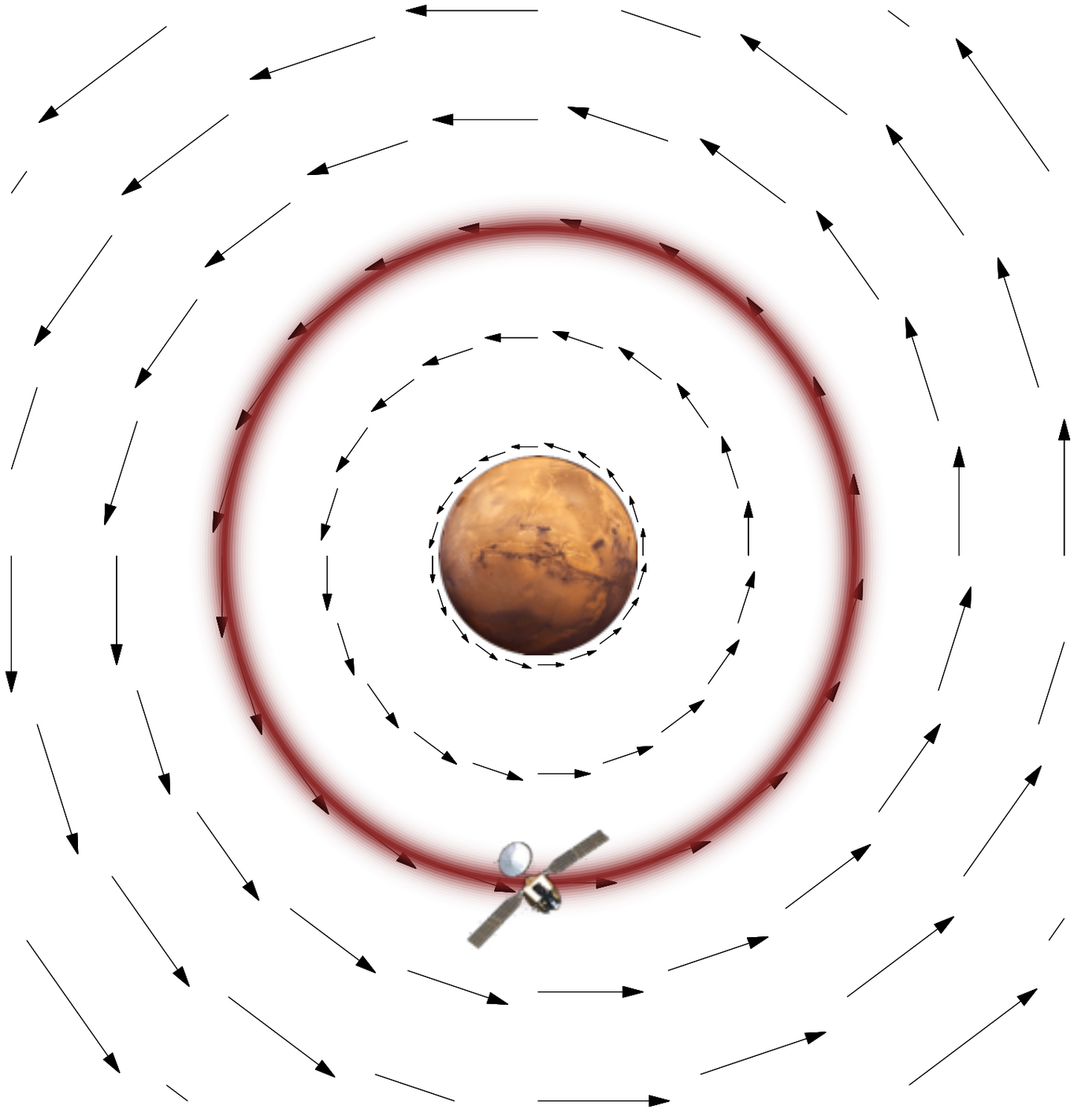
Too much and we fly off to infinity.



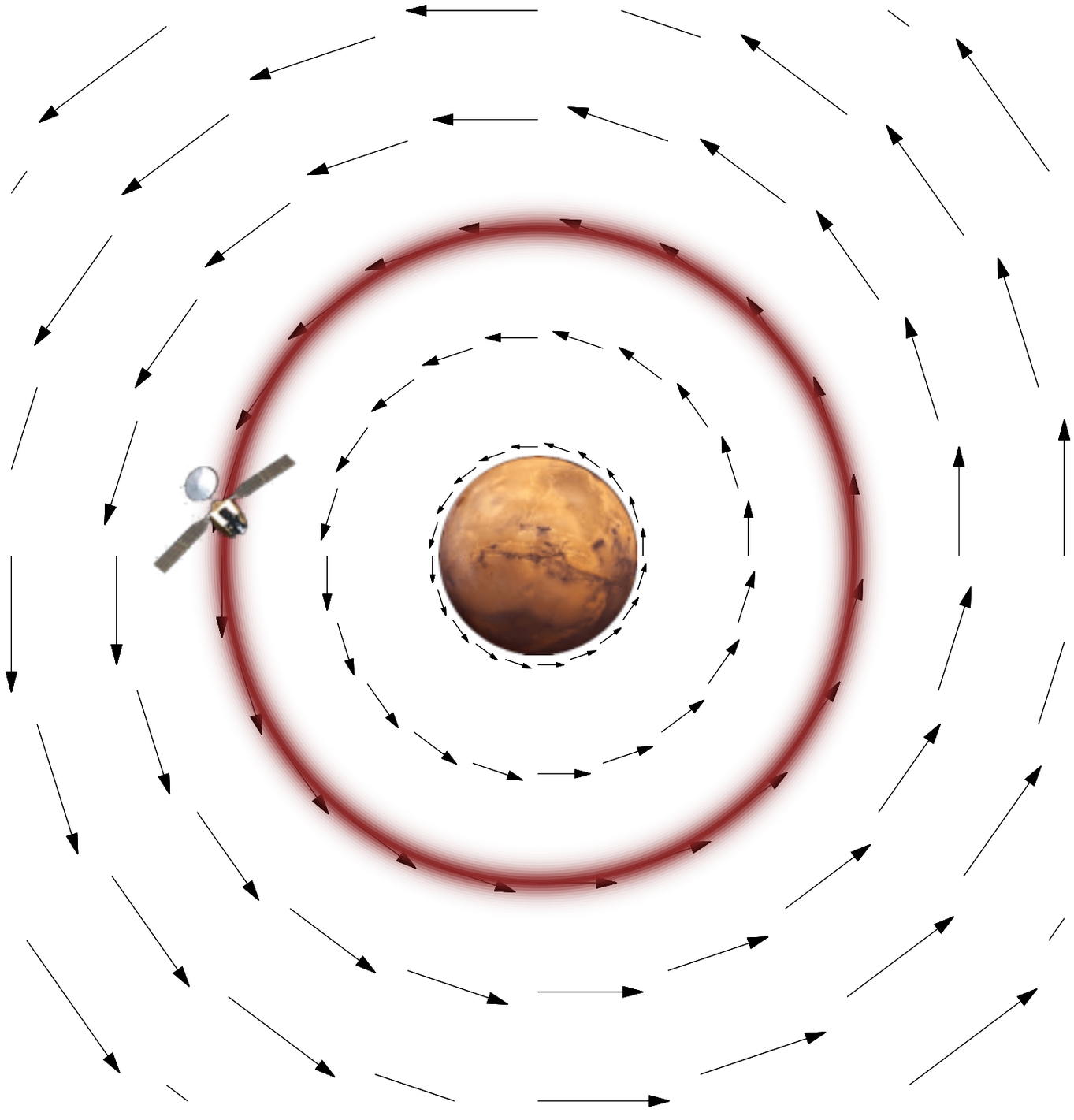
Too much and we fly off to infinity.

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

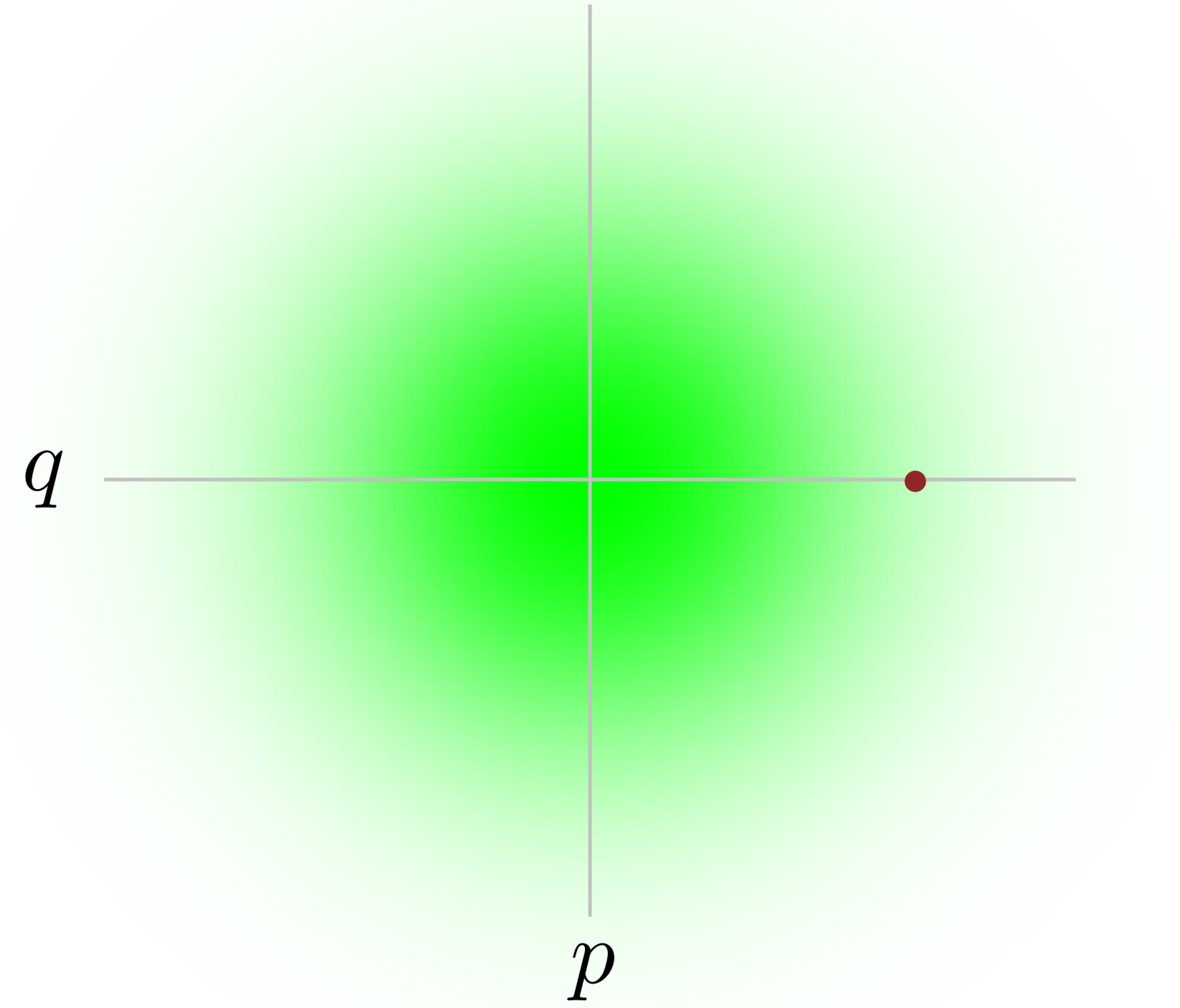
Just enough, however, aligns the gradients with the typical set and yields the desired orbital trajectory.



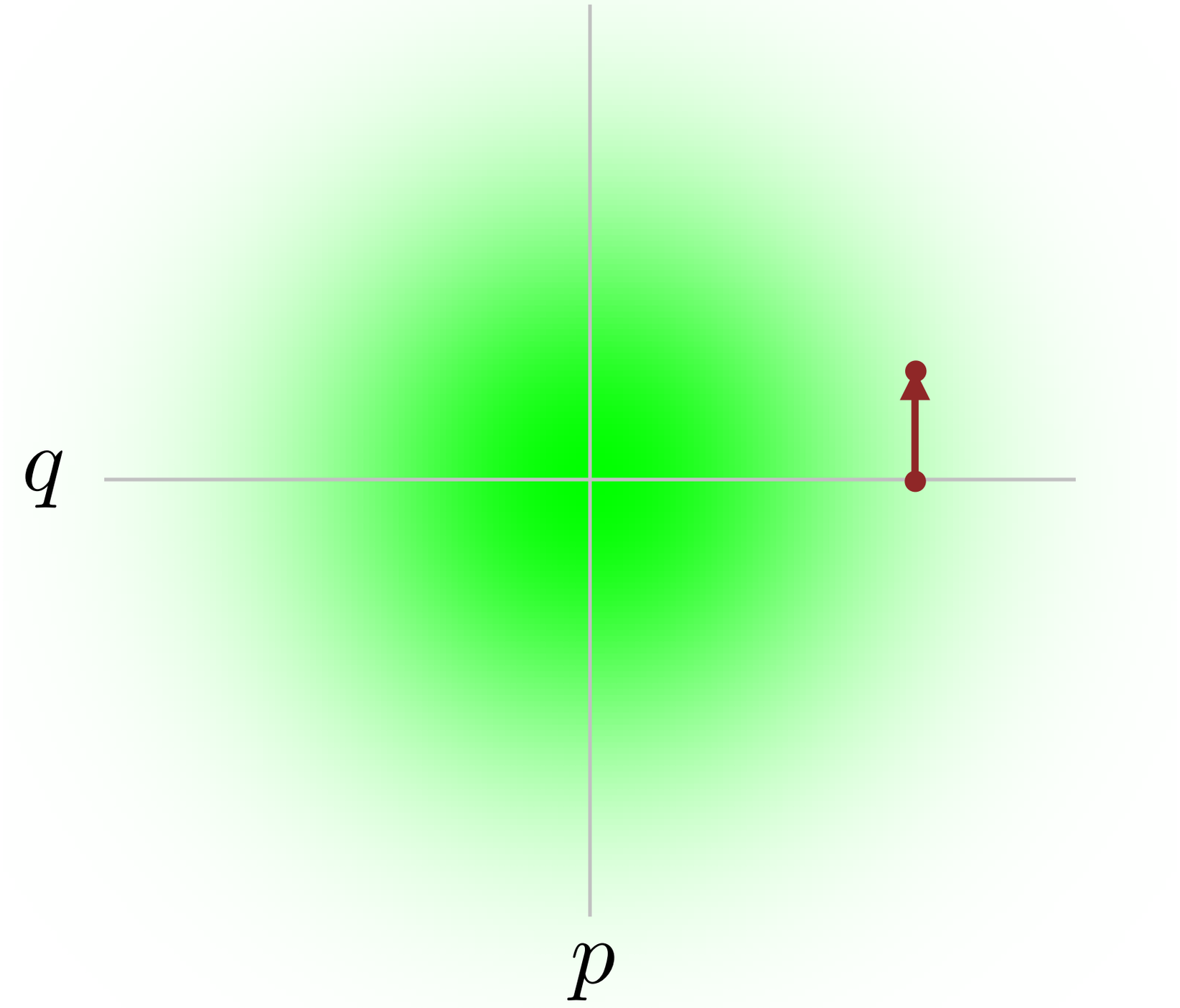
Just enough, however, aligns the gradients with the typical set and yields the desired orbital trajectory.



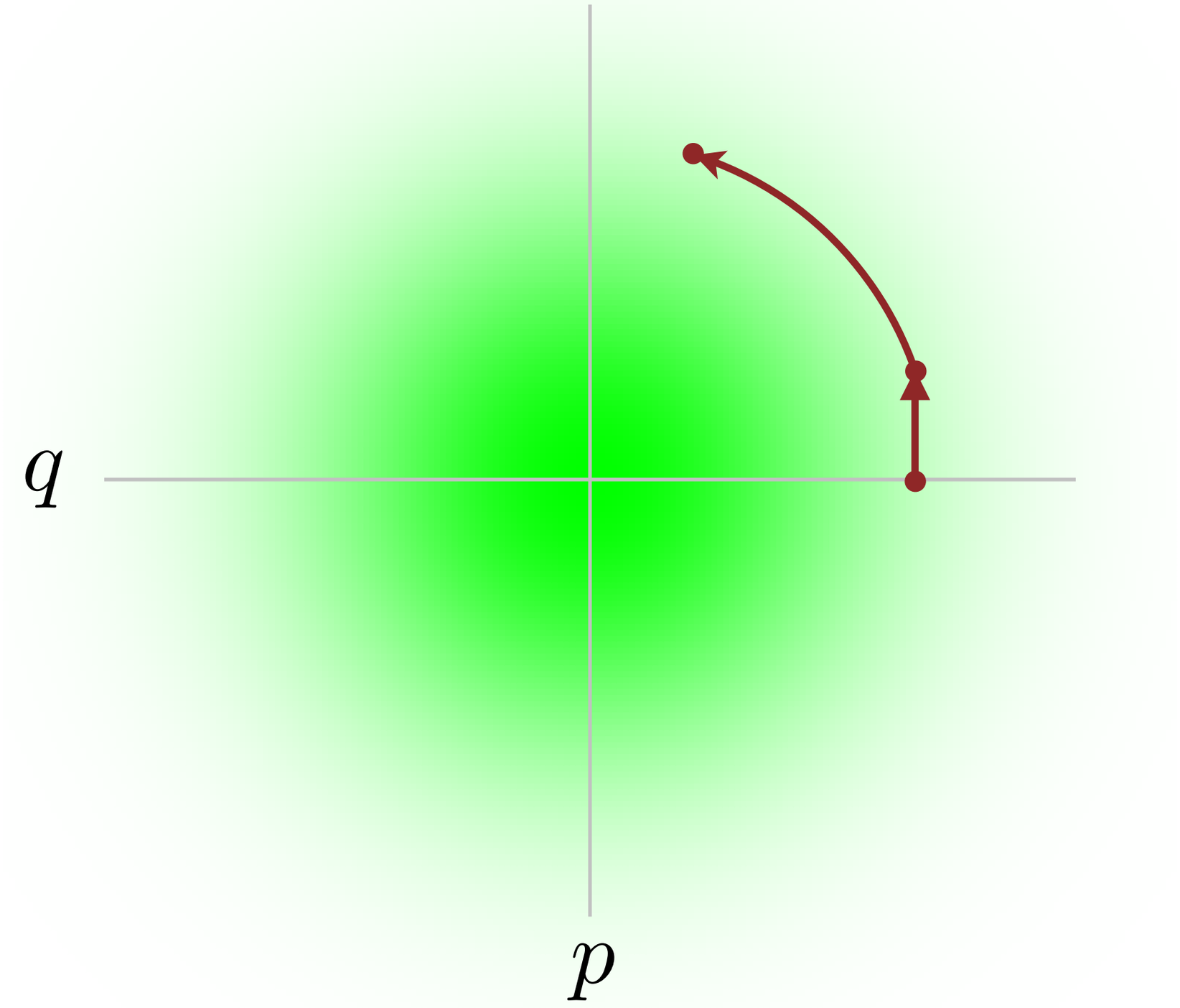
A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.



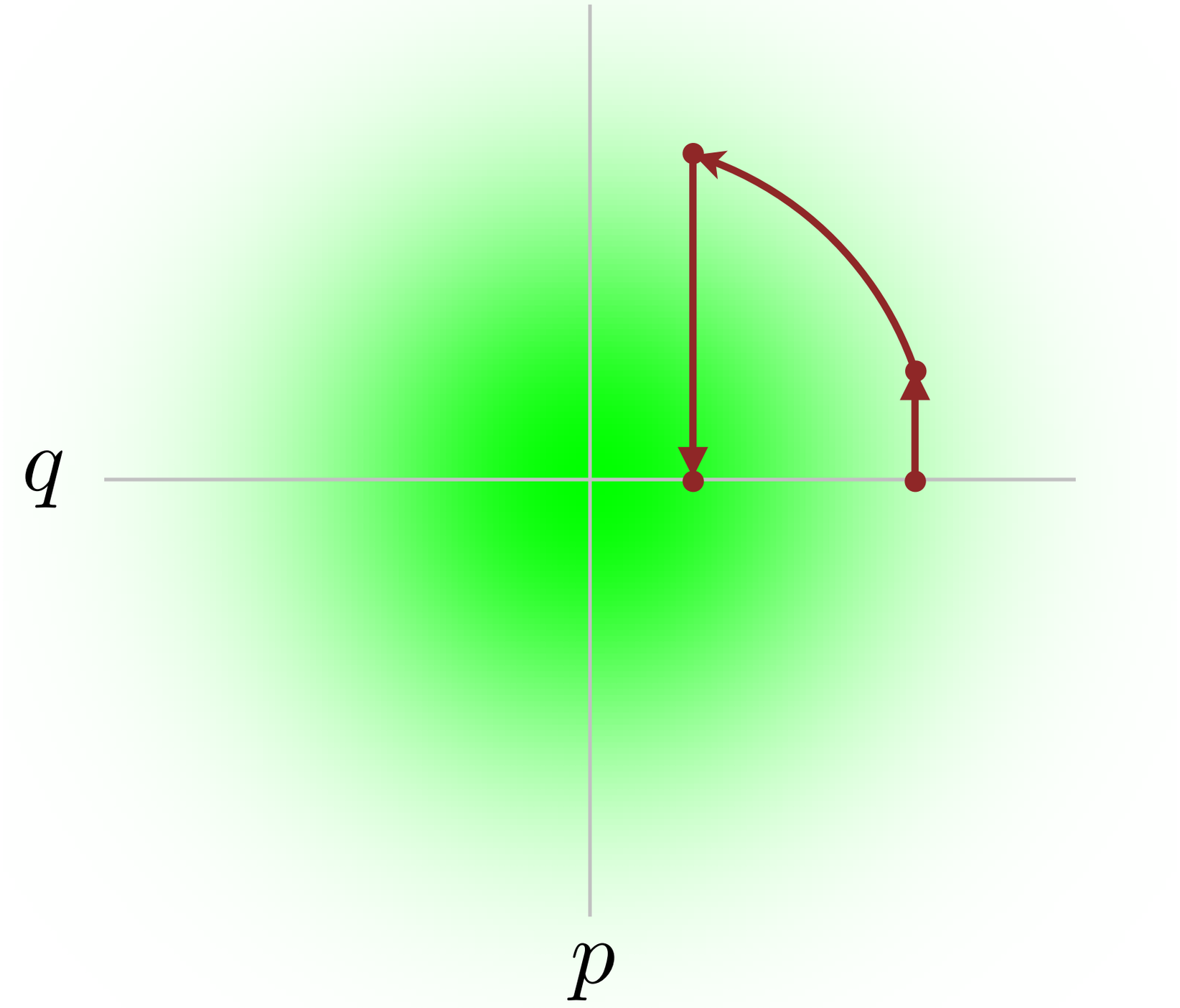
A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.



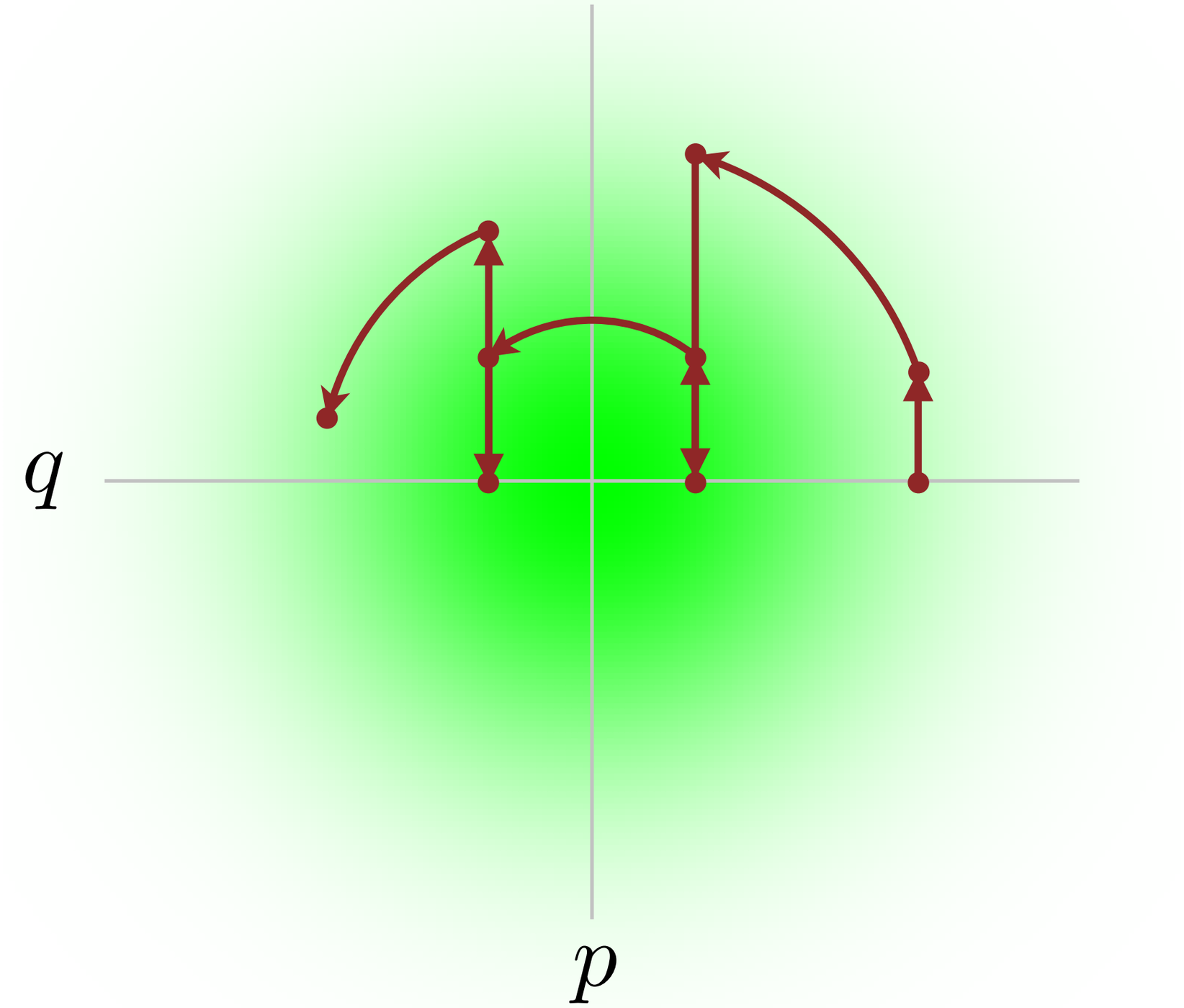
A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.



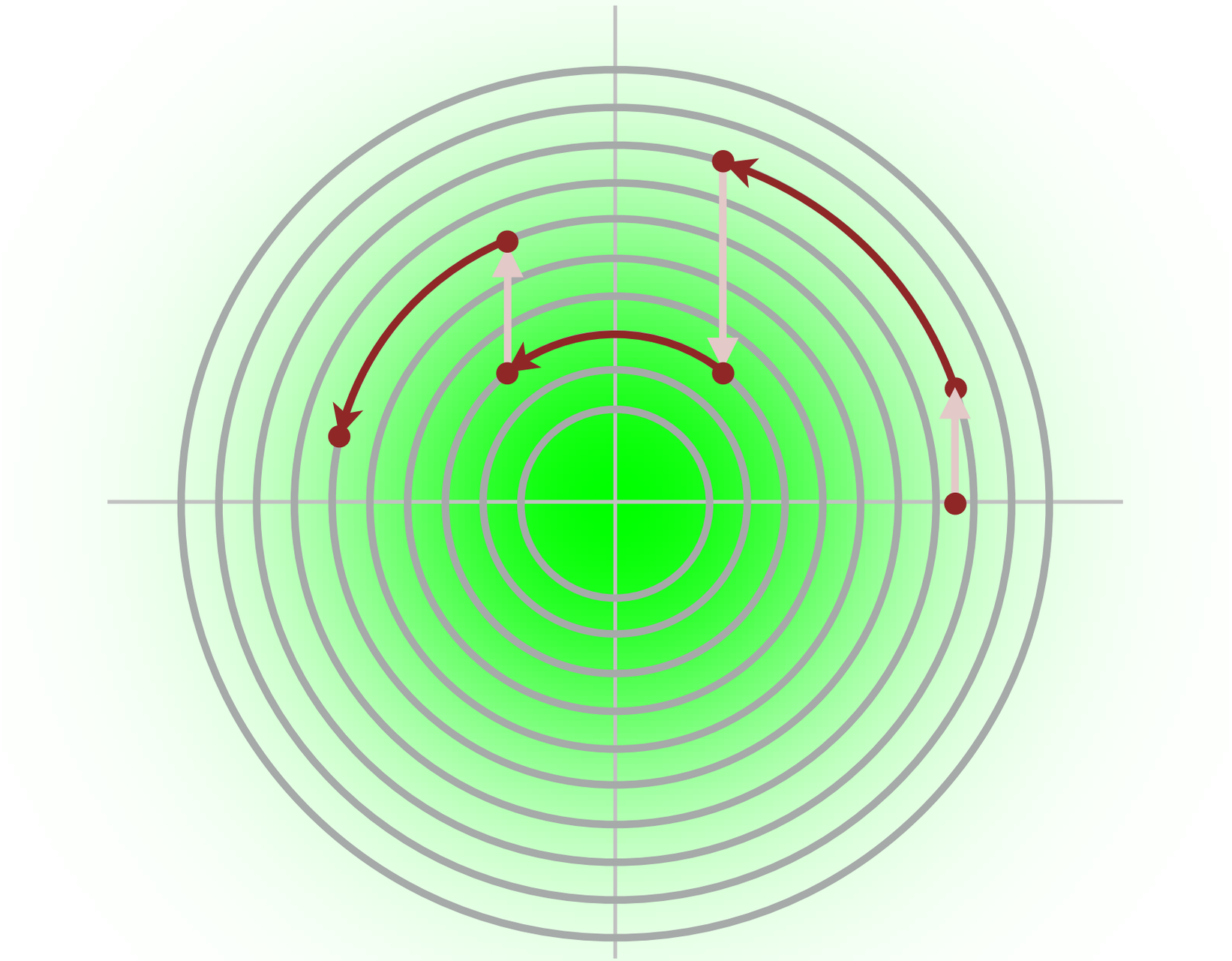
A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.



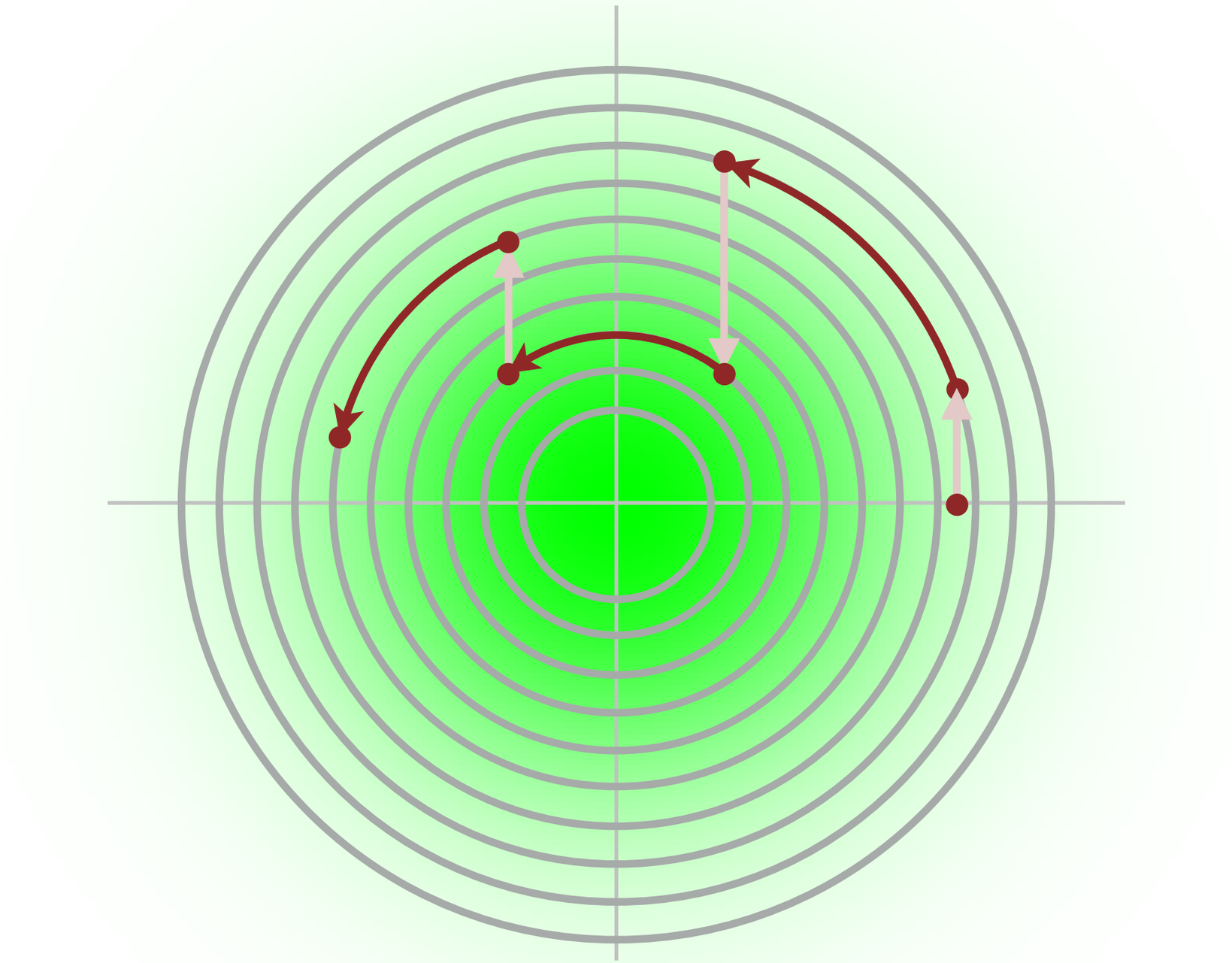
A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.



A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.

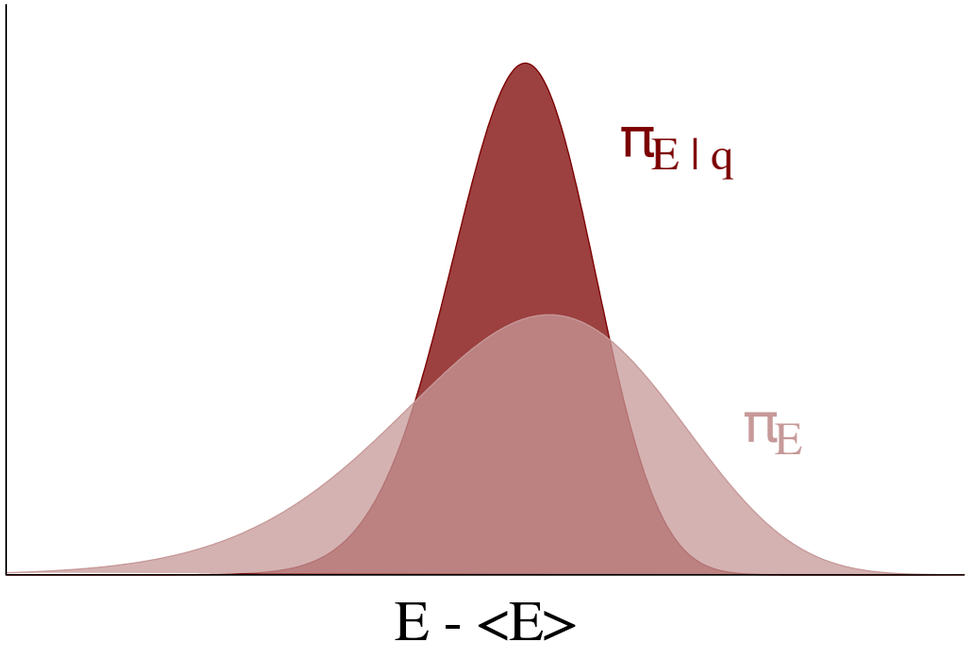


A rigorous theoretical foundation motivates not just the algorithm but also its optimized implementations.



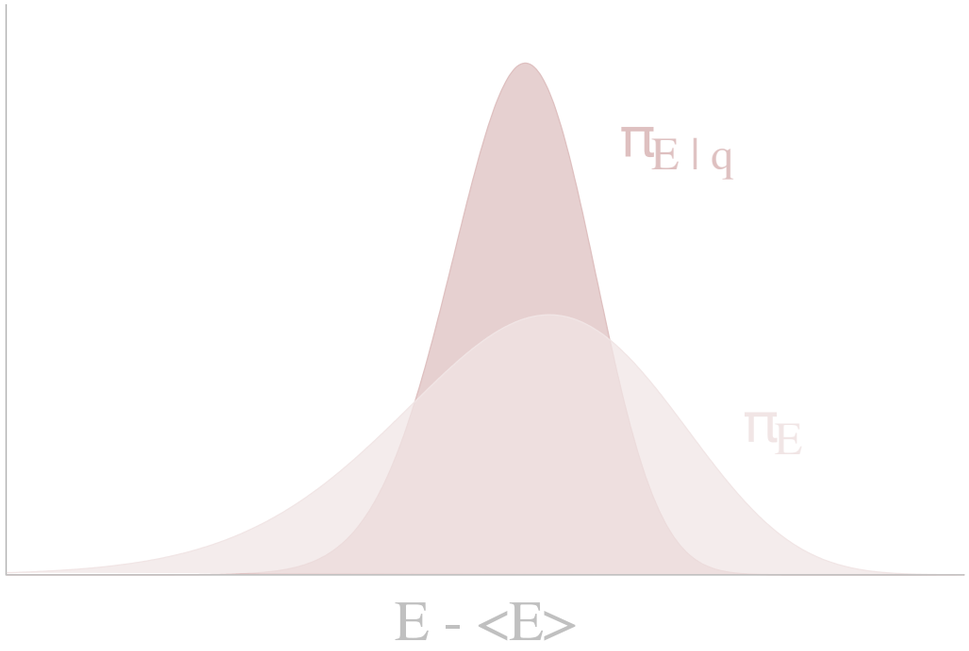
$$\pi_H = \pi_{H^{-1}(E)} \wedge \pi_E$$

The efficacy of the marginal energy exploration provides a criterion for optimizing the kinetic energy.

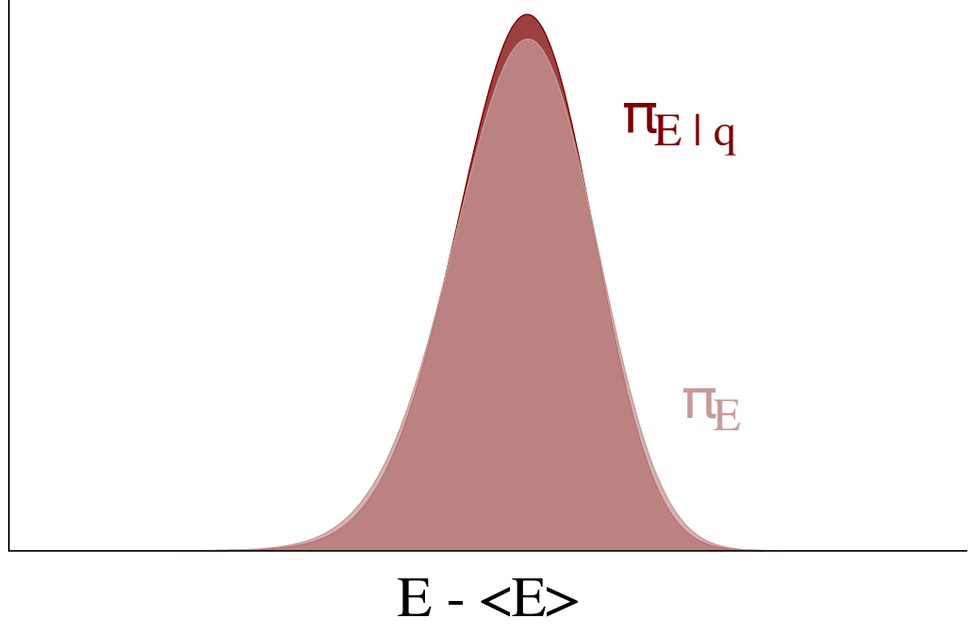


Large Autocorrelations

The efficacy of the marginal energy exploration provides a criterion for optimizing the kinetic energy.

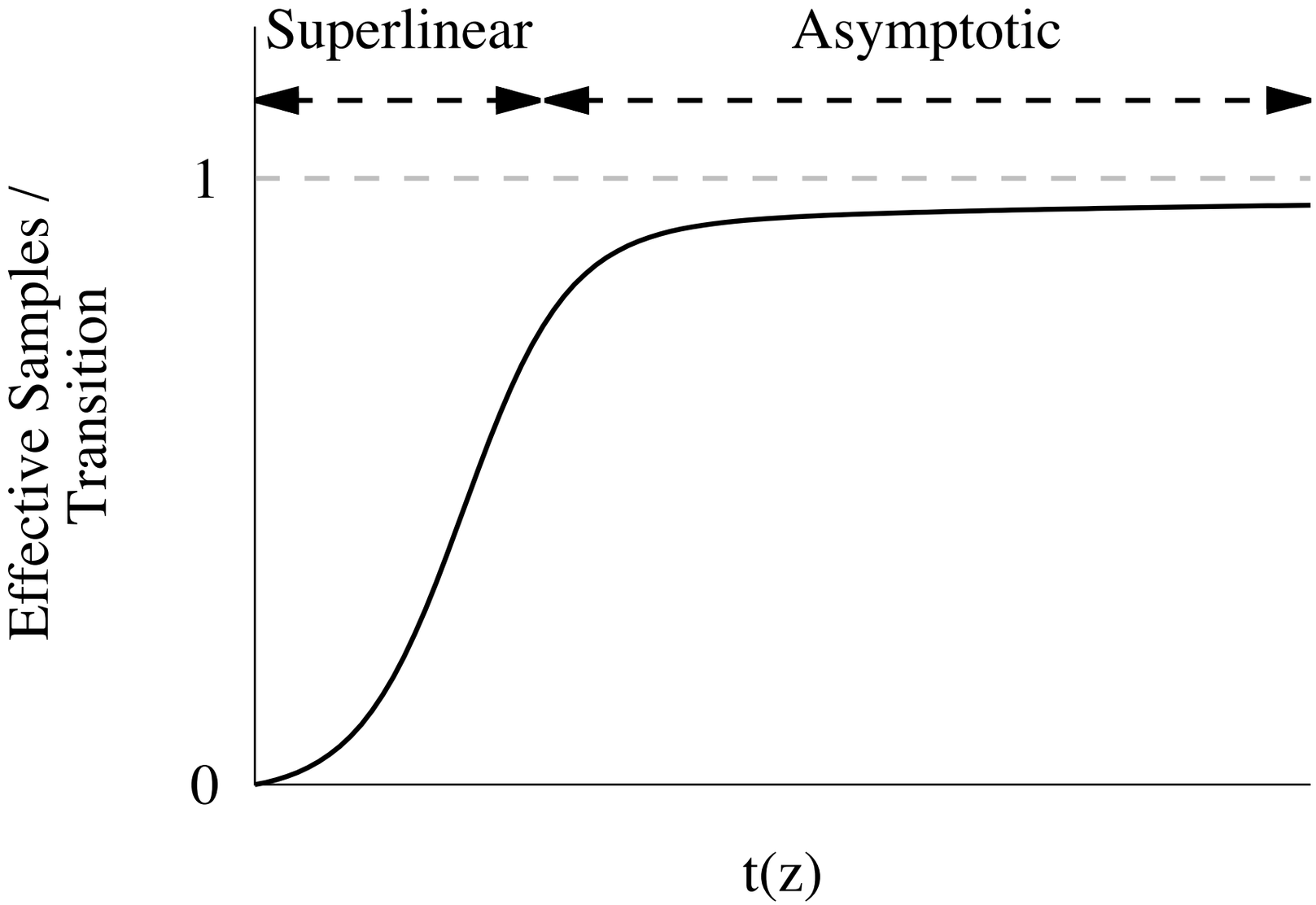


Large Autocorrelations

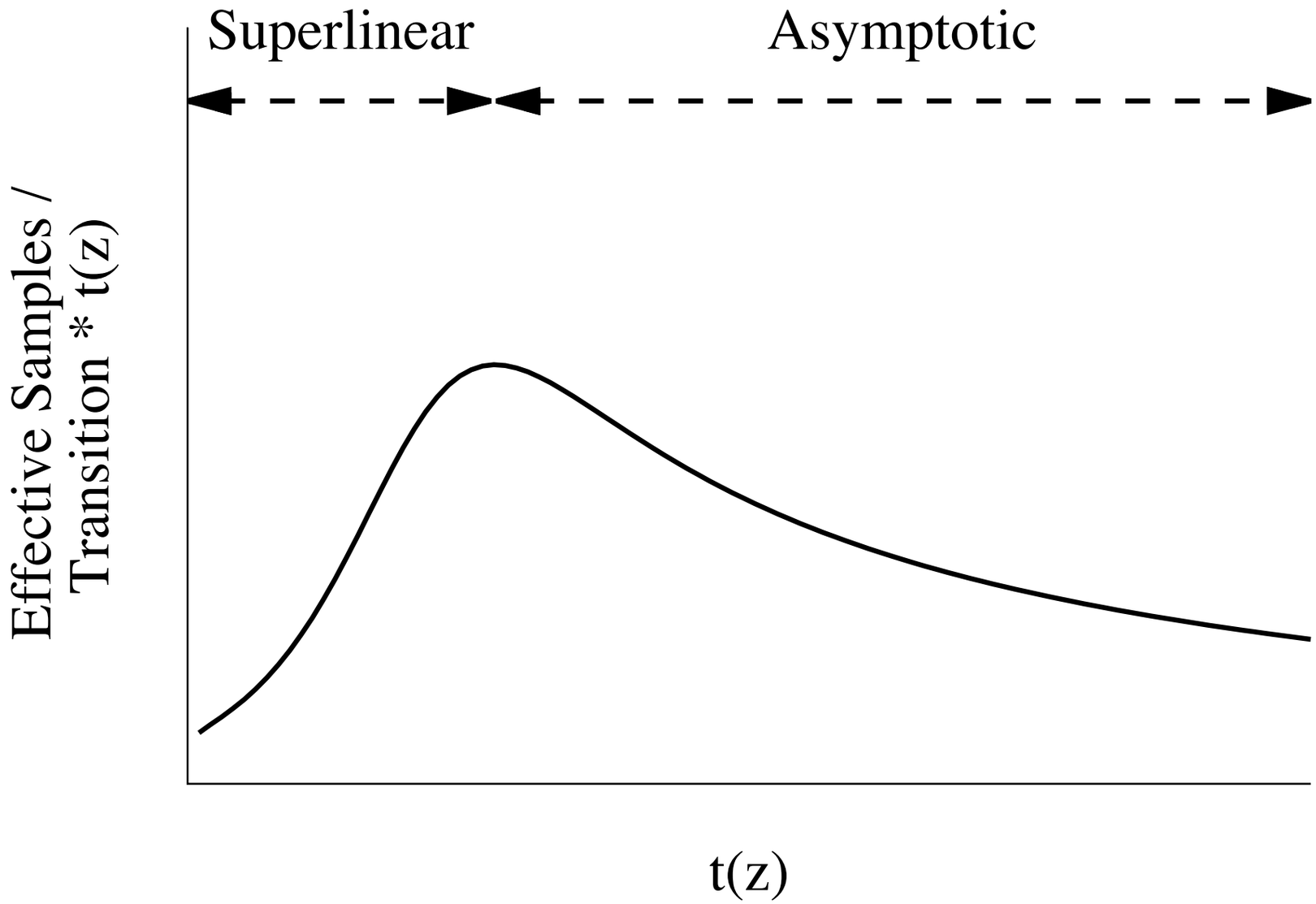


Small Autocorrelations

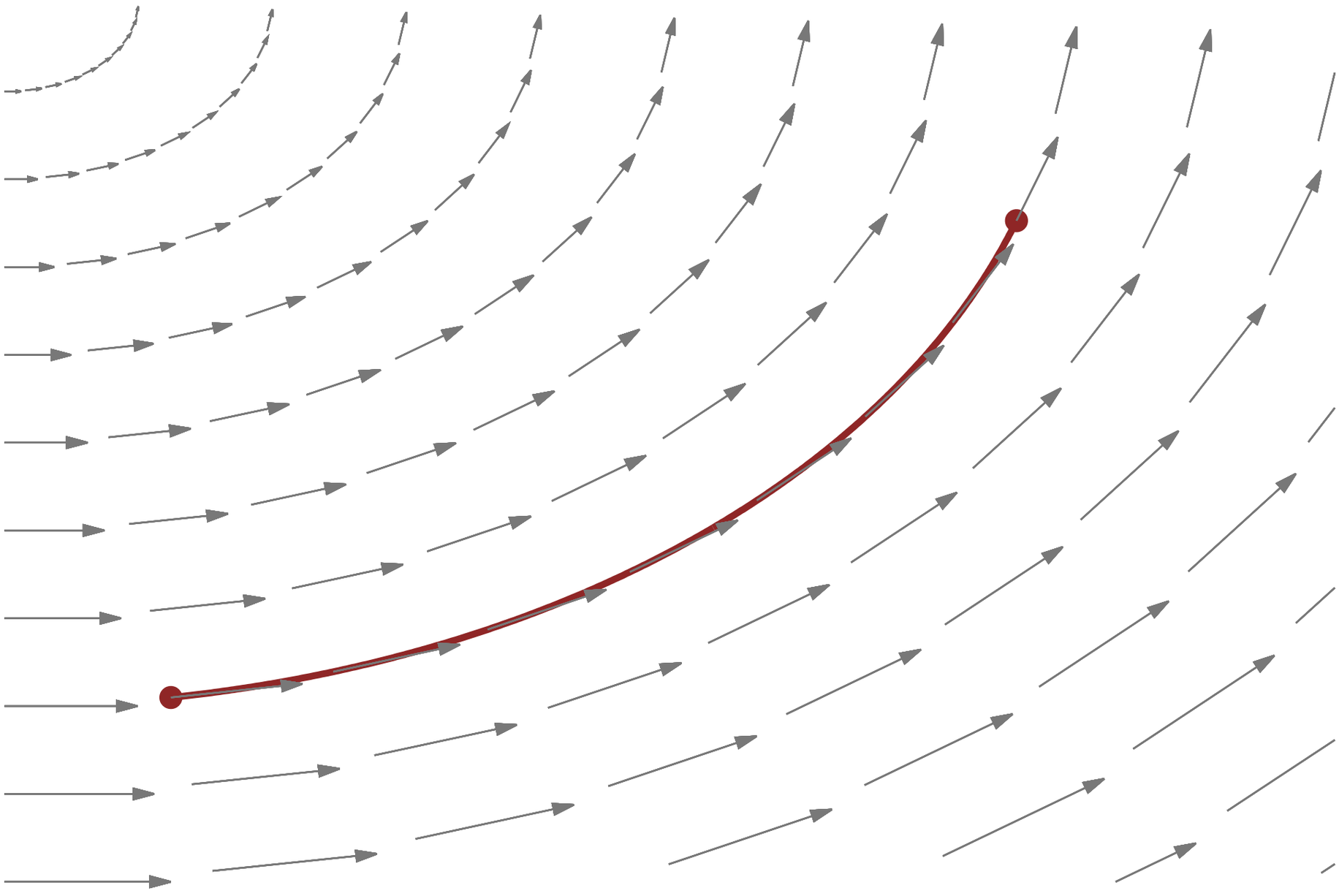
Similarly, the efficacy of the Hamiltonian flow provides a criterion for dynamically tuning the integration time.



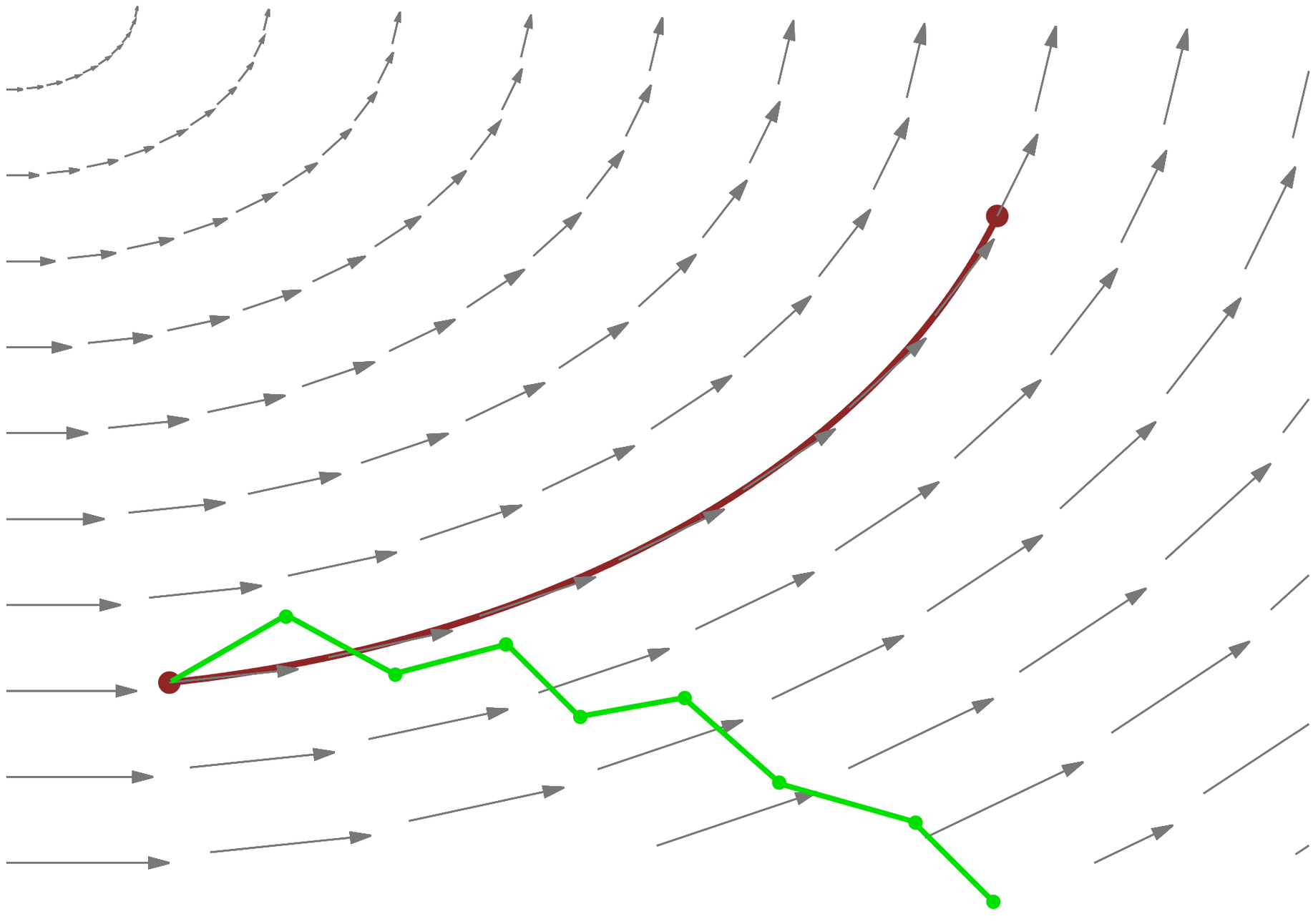
Similarly, the efficacy of the Hamiltonian flow provides a criterion for dynamically tuning the integration time.



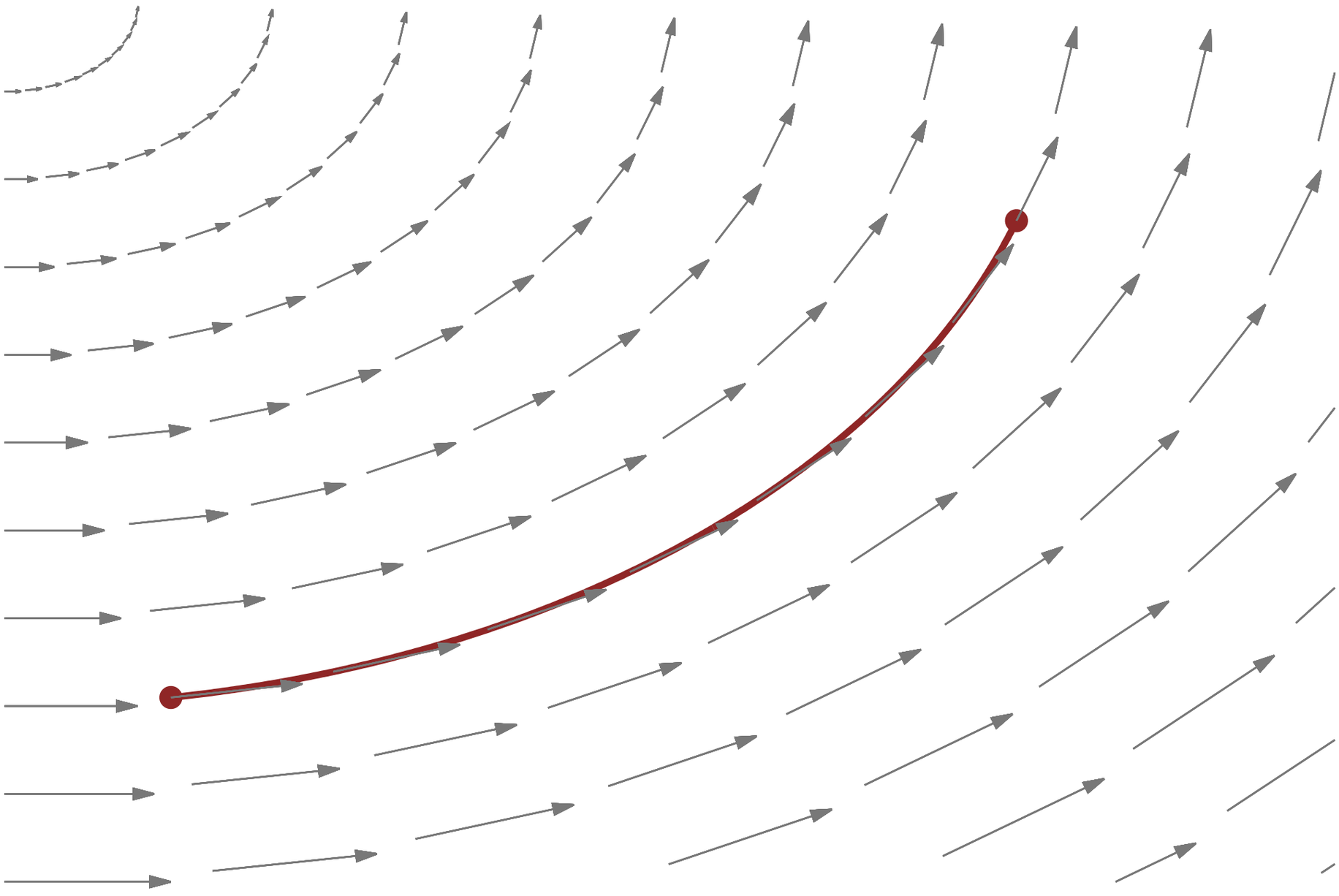
The microcanonical perspective also motivates scalable numerical implementations using *symplectic integrators*.



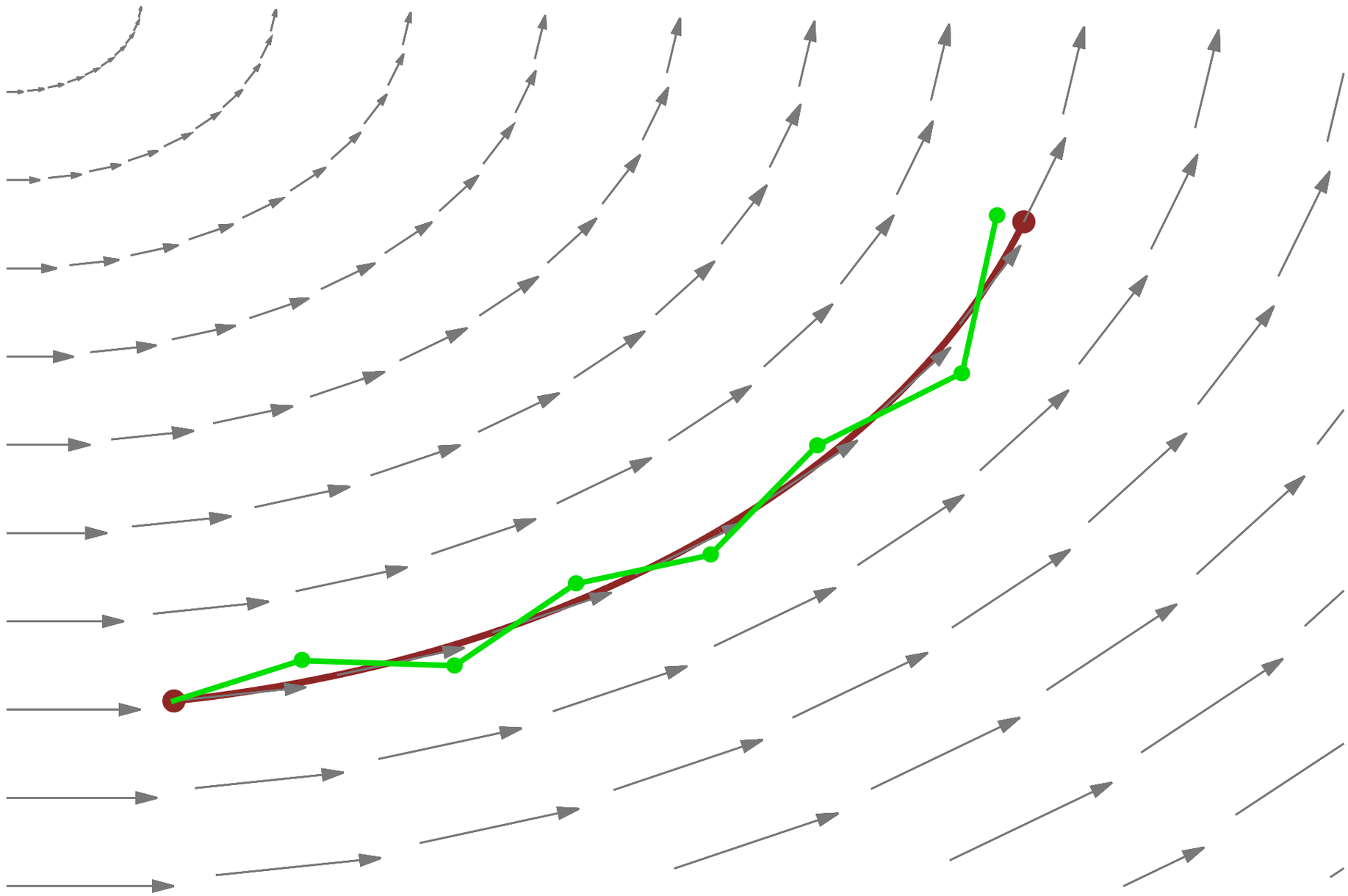
The microcanonical perspective also motivates scalable numerical implementations using *symplectic integrators*.



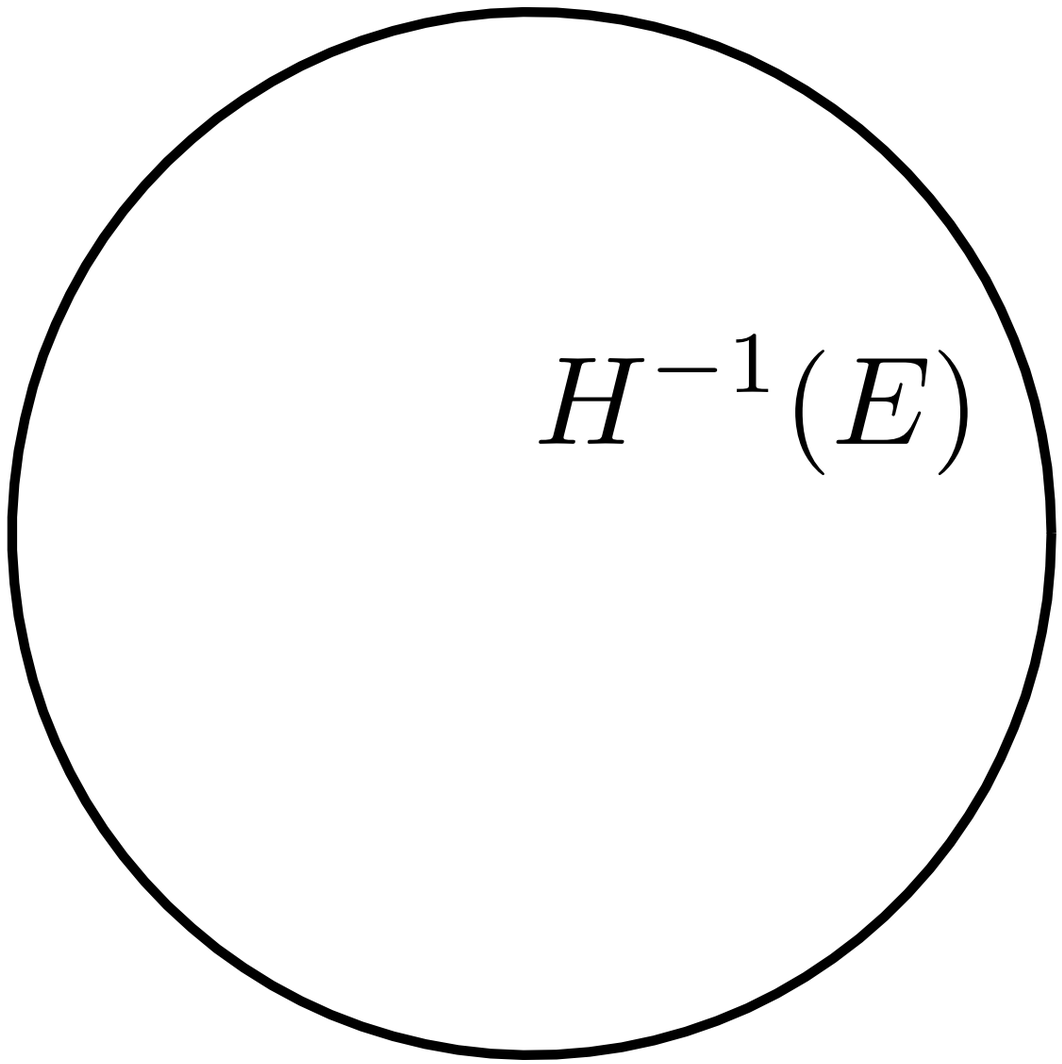
The microcanonical perspective also motivates scalable numerical implementations using *symplectic integrators*.



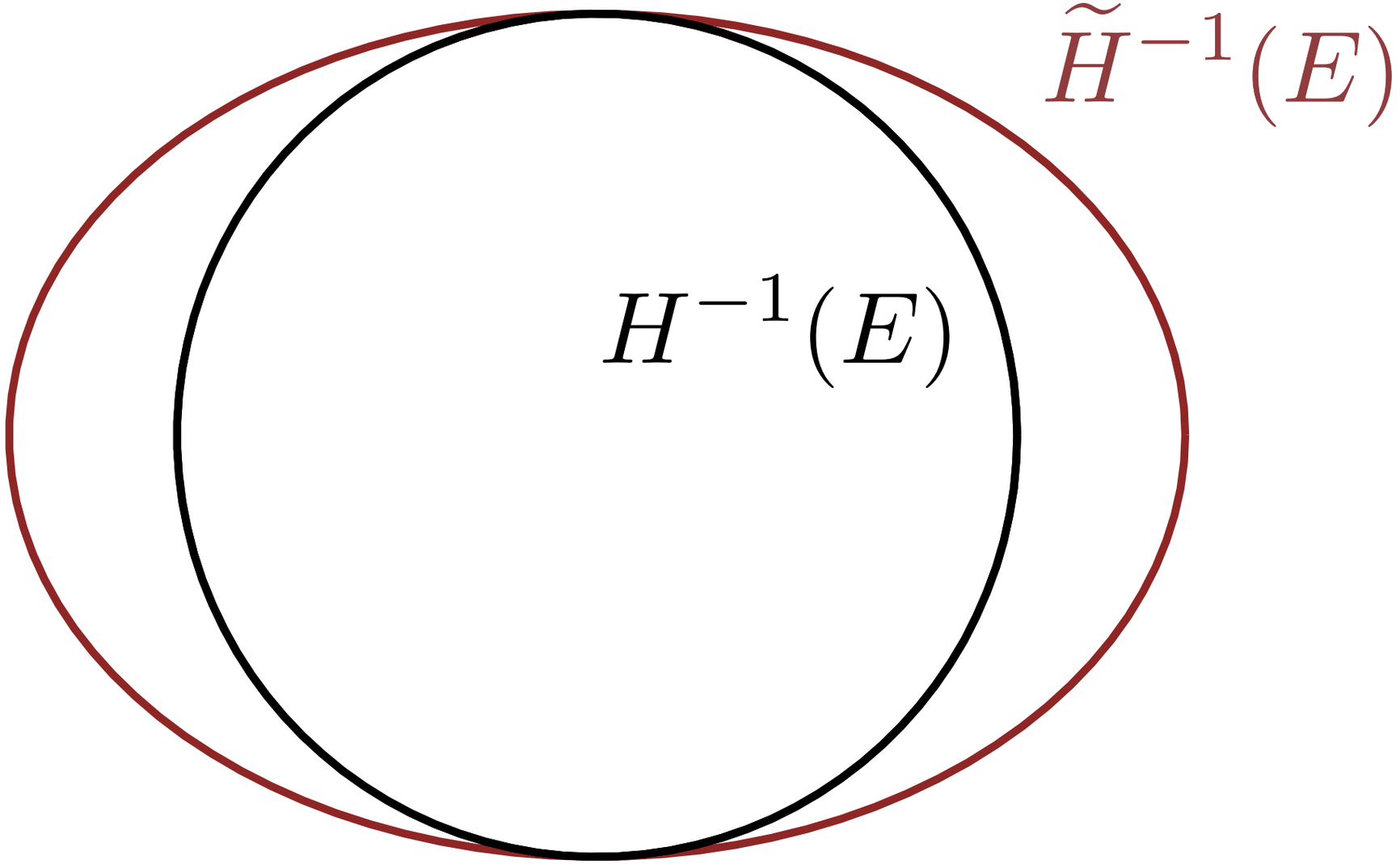
The microcanonical perspective also motivates scalable numerical implementations using *symplectic integrators*.



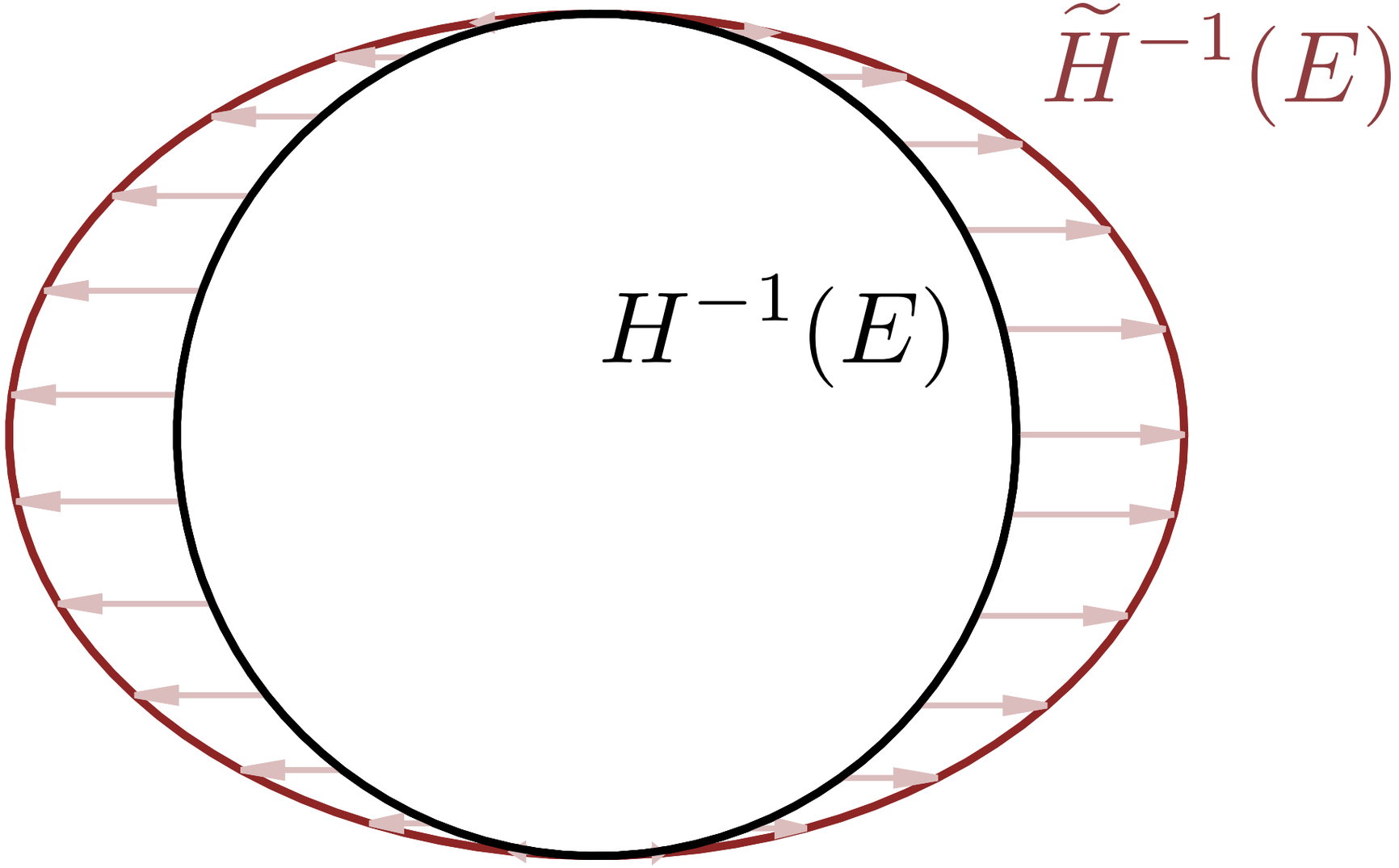
Finally, the interaction of the symplectic integrator and the microcanonical geometry motivates optimal tuning.



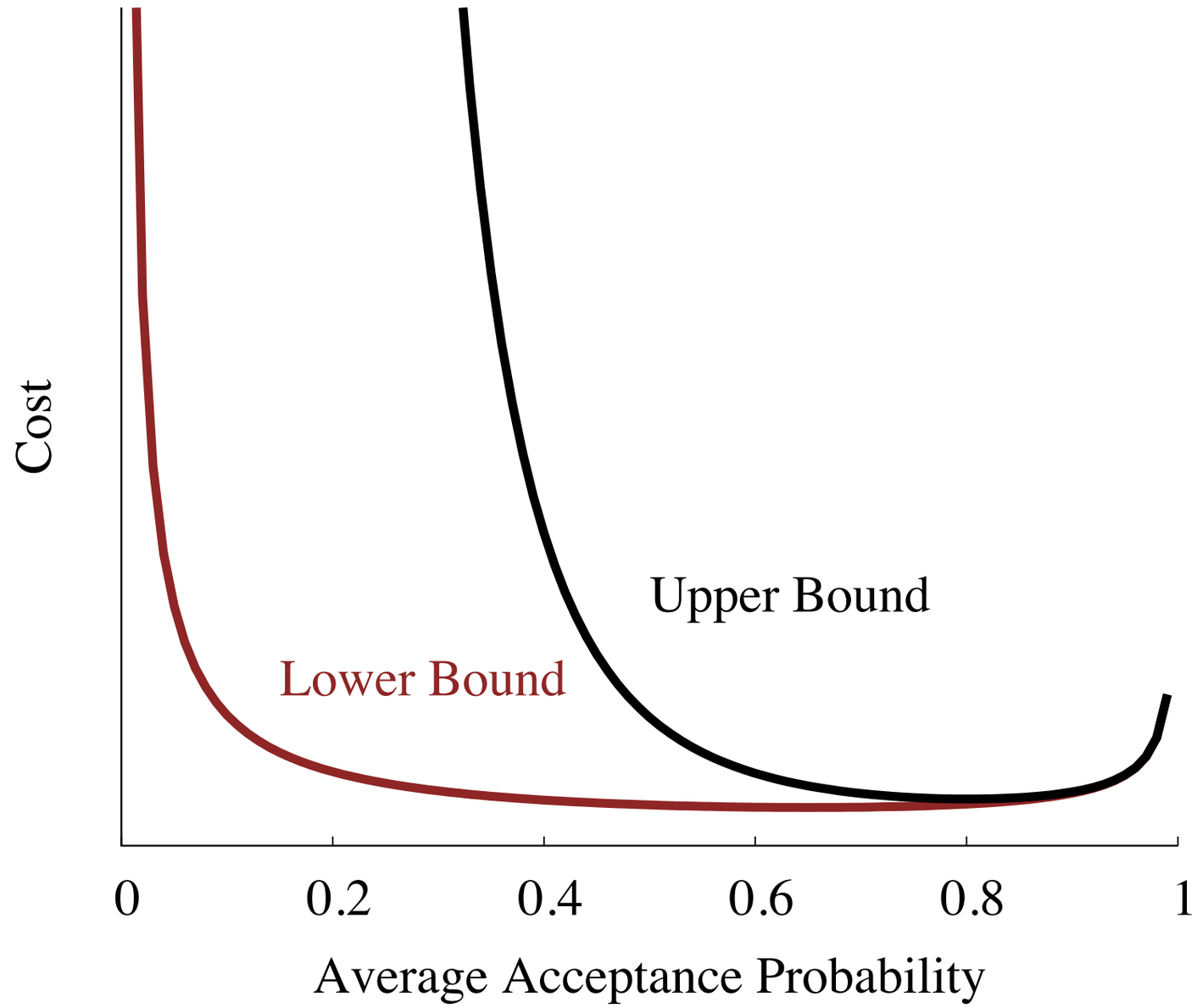
Finally, the interaction of the symplectic integrator and the microcanonical geometry motivates optimal tuning.



Finally, the interaction of the symplectic integrator and the microcanonical geometry motivates optimal tuning.



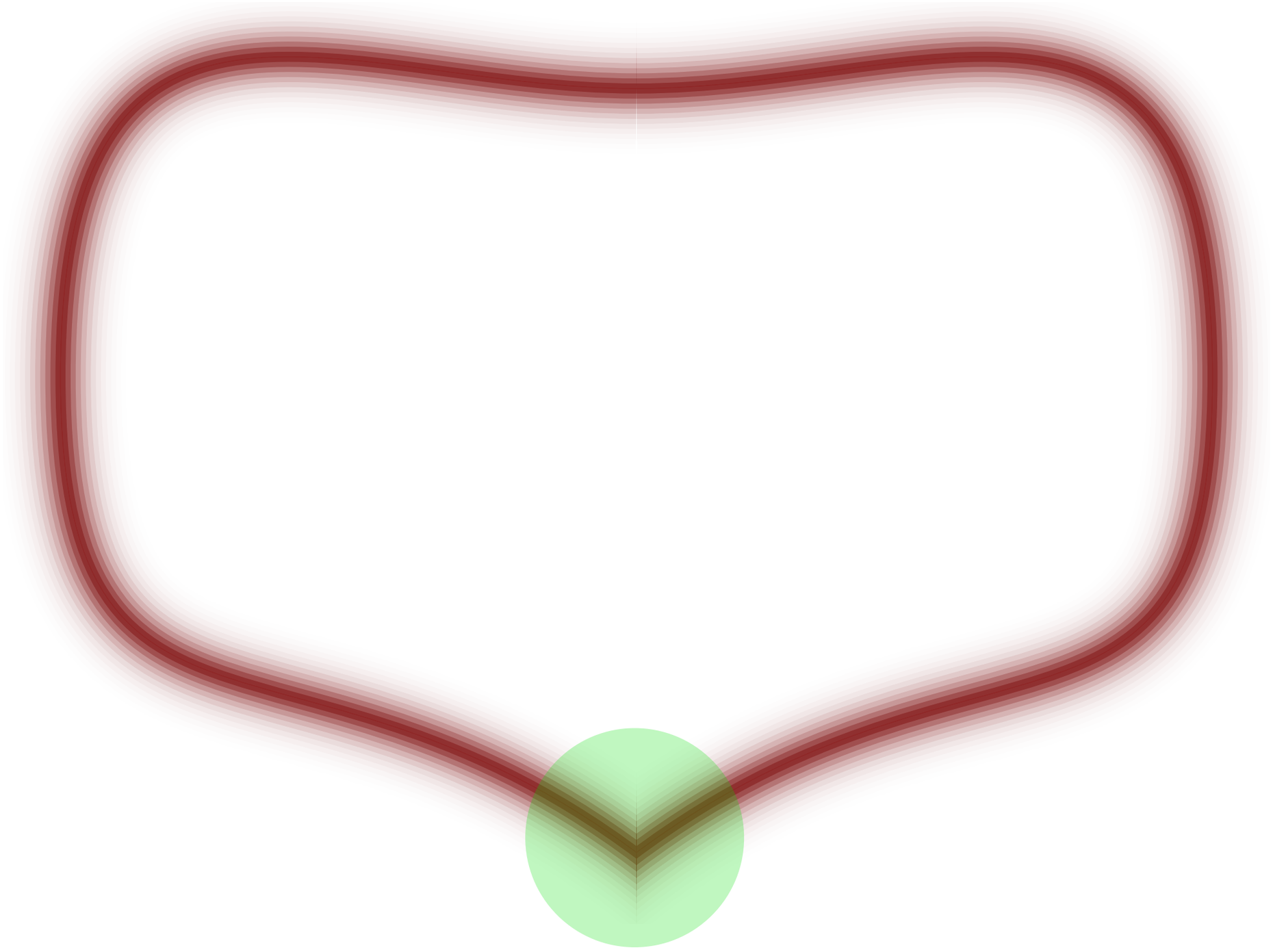
Finally, the interaction of the symplectic integrator and the microcanonical geometry motivates optimal tuning.



We can achieve a computationally efficient implementation, but how robust will it be?

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} f(q_n) \to \mathbb{E}_\pi[f]$$

We can achieve a computationally efficient implementation, but how robust will it be?



To ensure productive behavior after only finite iterations we need to verify conditions like *geometric ergodicity*.

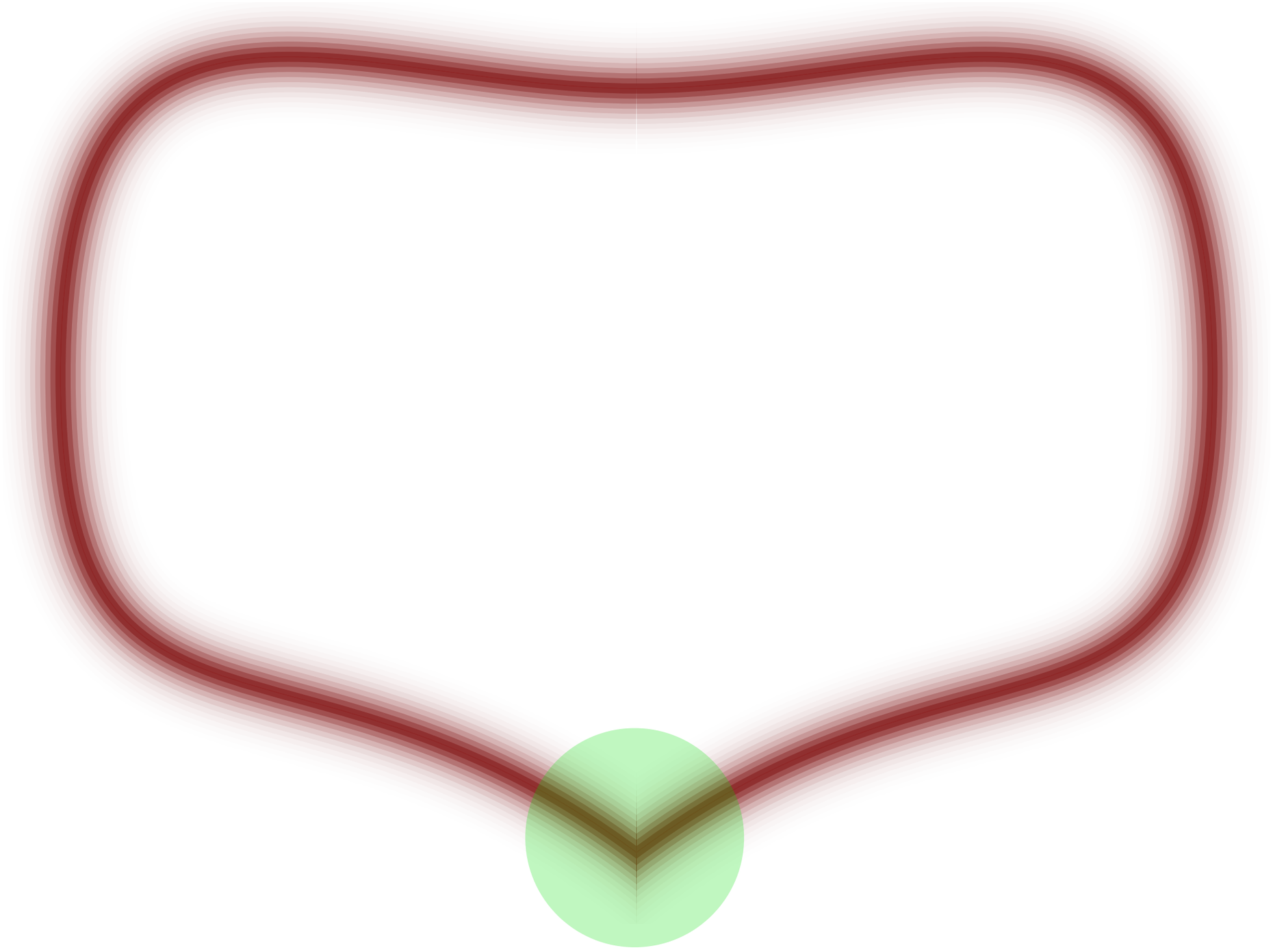
$$\|\mathcal{T}^n \delta_{q_0} - \pi\| \leq C\rho^n$$

To ensure productive behavior after only finite iterations we need to verify conditions like *geometric ergodicity*.

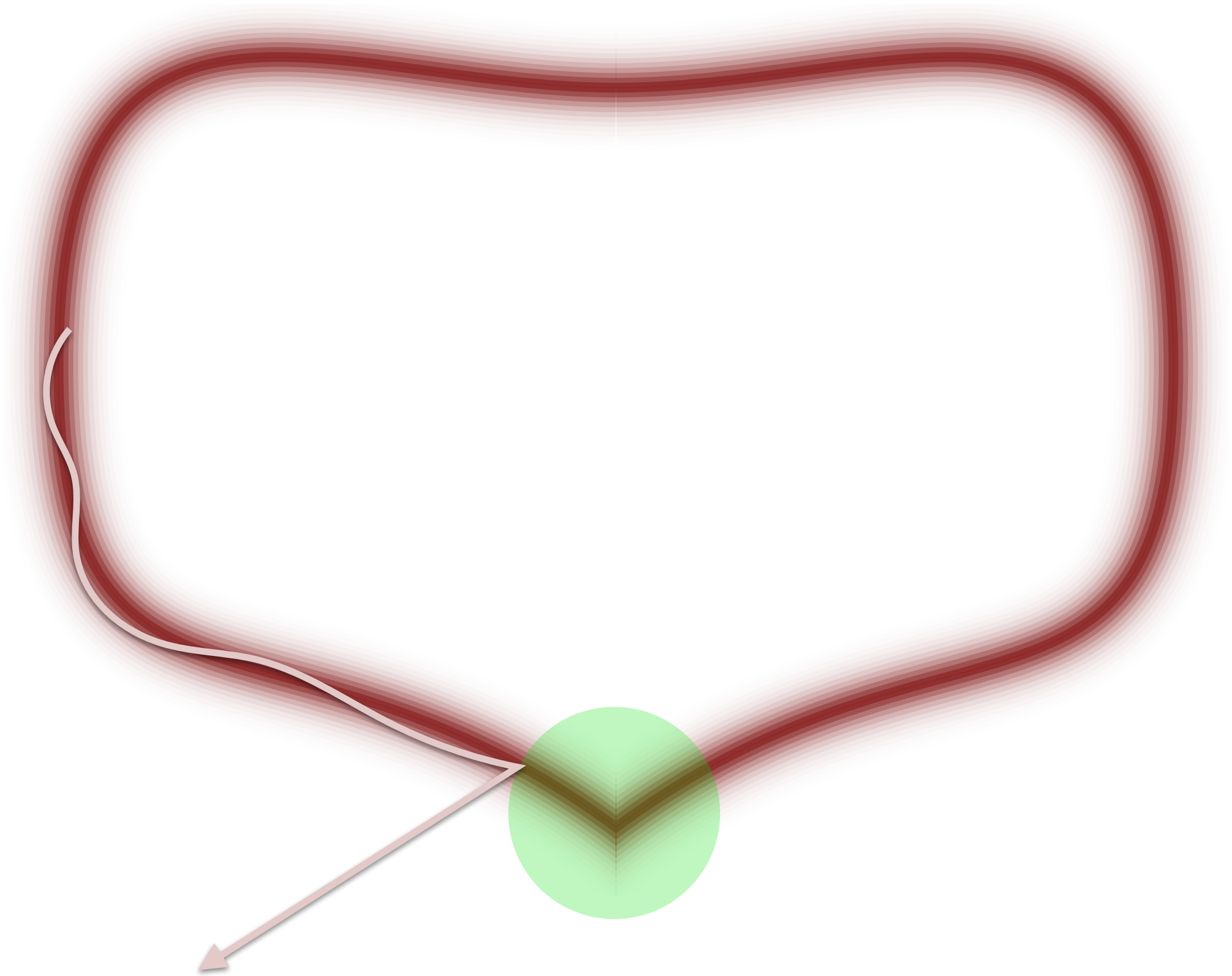
$$\|\mathcal{T}^n \delta_{q_0} - \pi\| \leq C\rho^n$$

$$\frac{\hat{f}_N(q_0) - \mathbb{E}_\pi[f]}{\sqrt{\lambda N}} \xrightarrow{\mathcal{D}} \mathcal{N}(0,1)$$

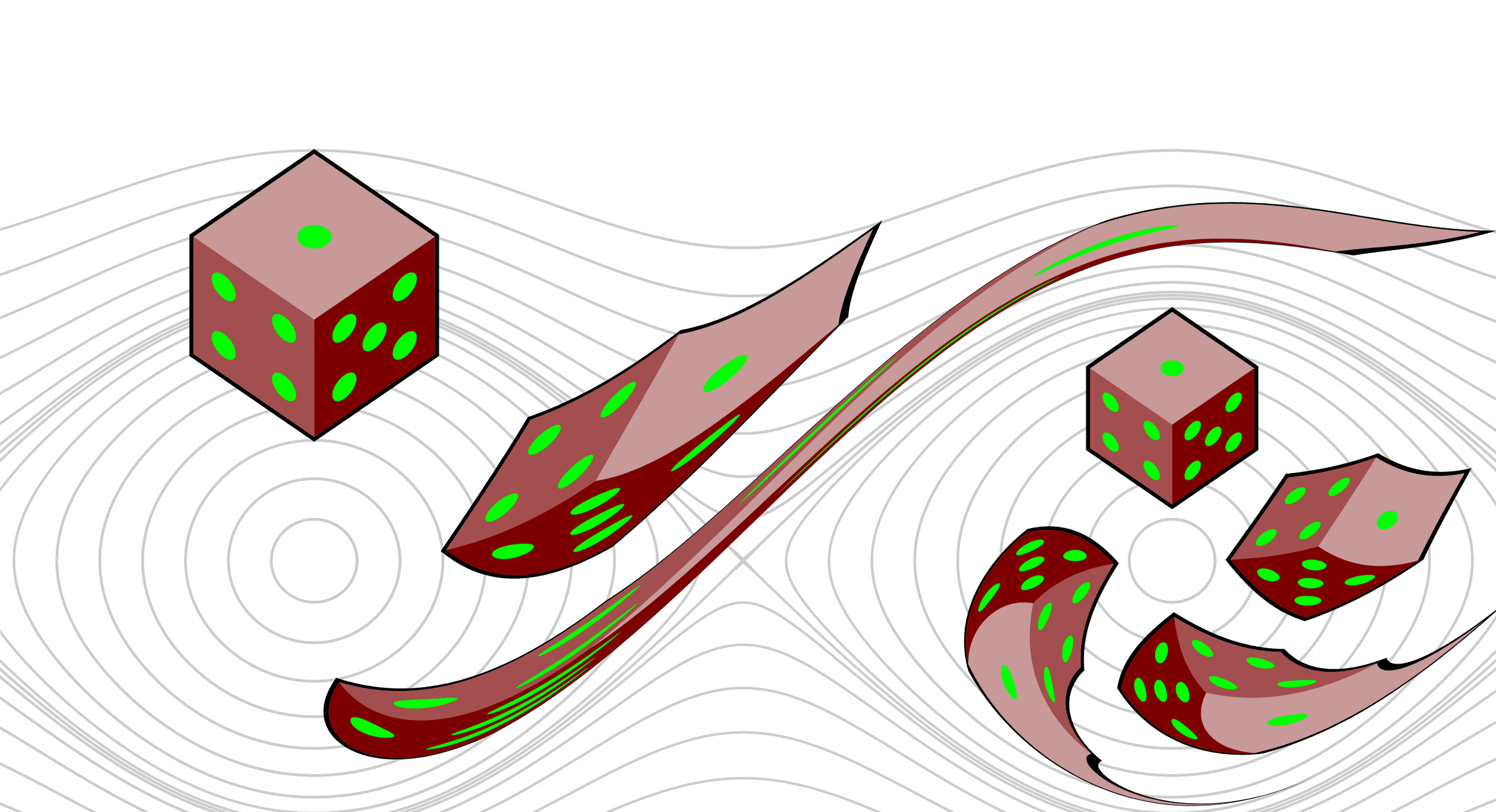
The geometry motivates techniques for proving geometry ergodicity in theory and diagnosing failures in practice.



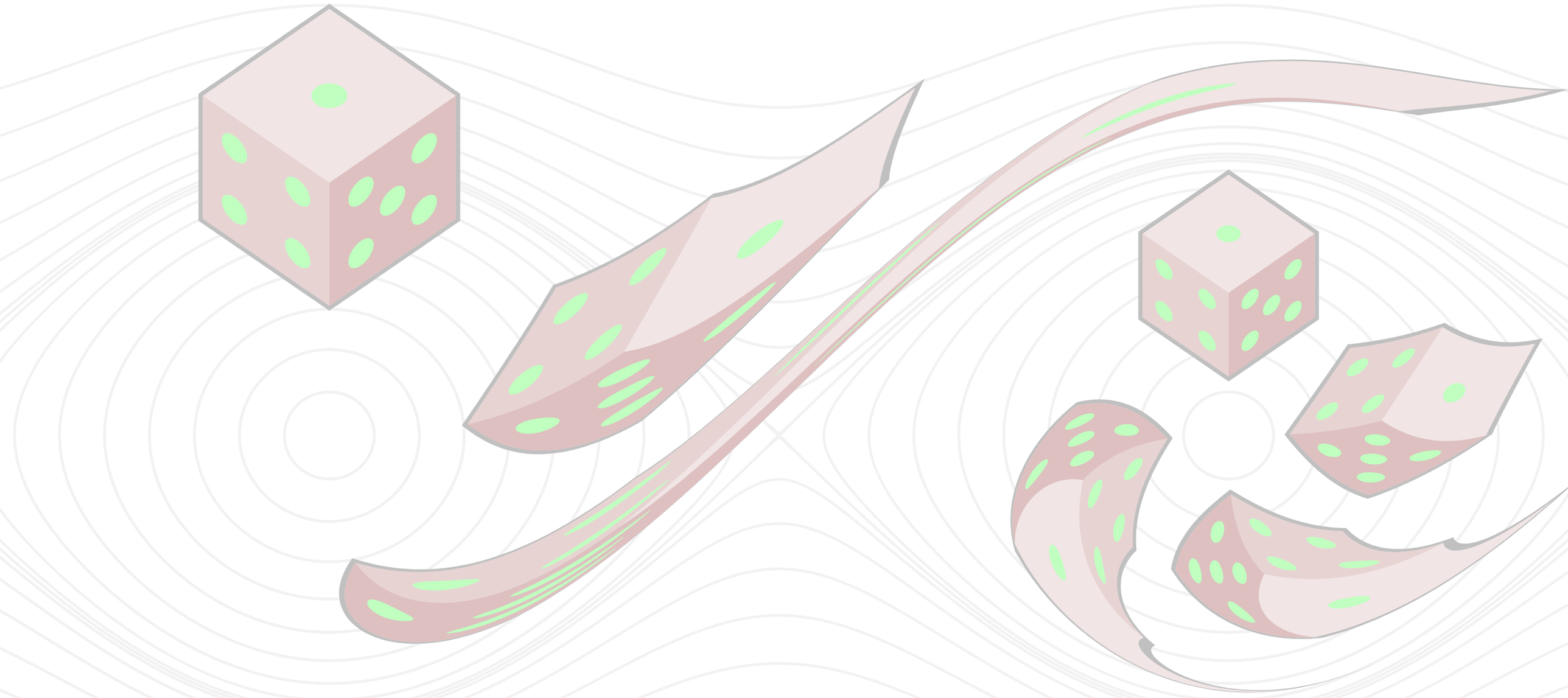
The geometry motivates techniques for proving geometry ergodicity in theory and diagnosing failures in practice.



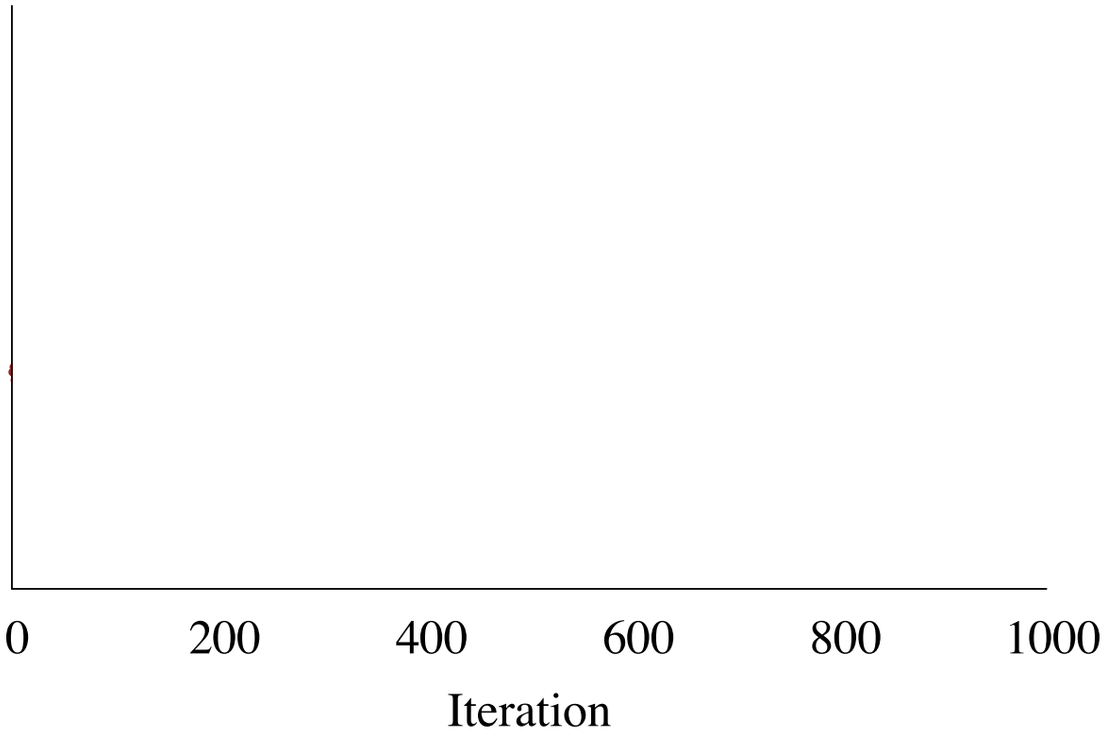
This understand can then be used to build high-performance implementations in tools like Stan.

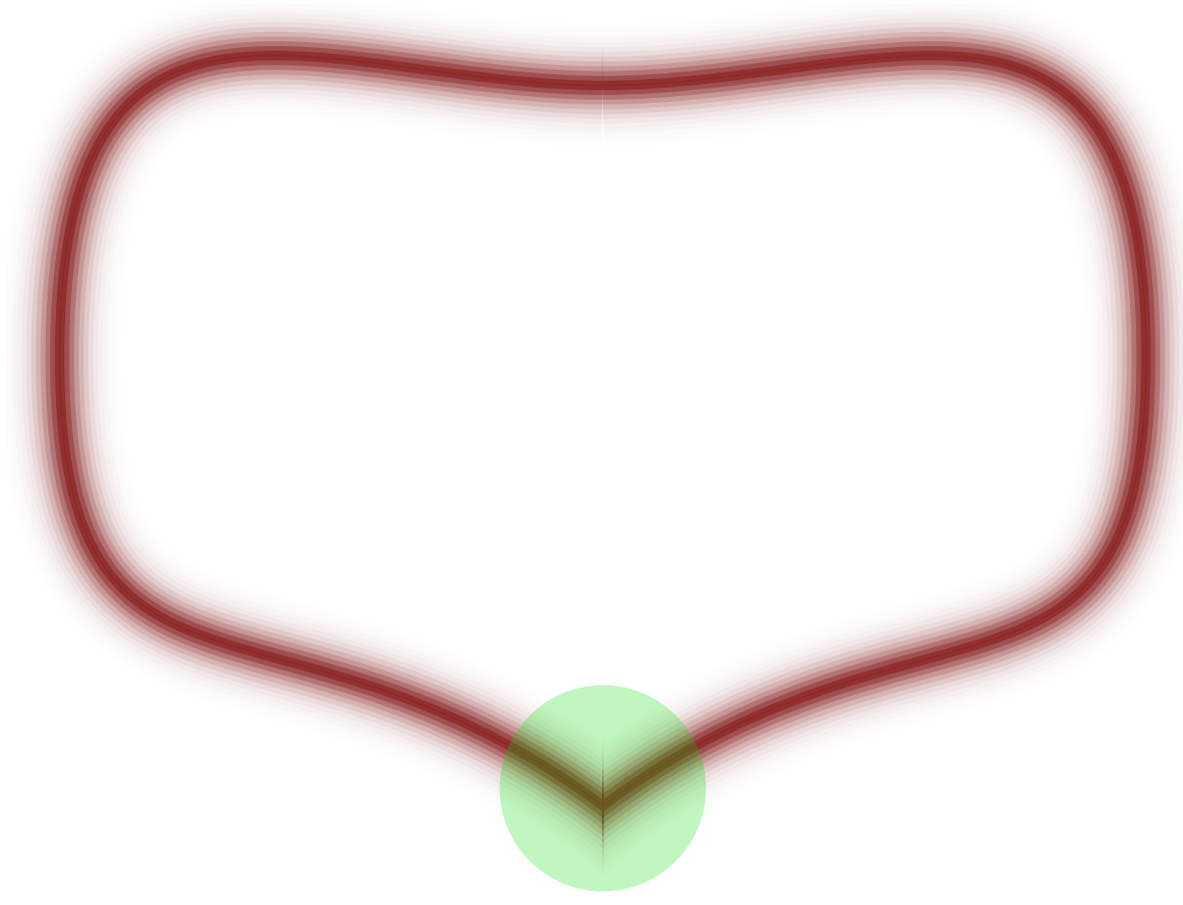


Backups

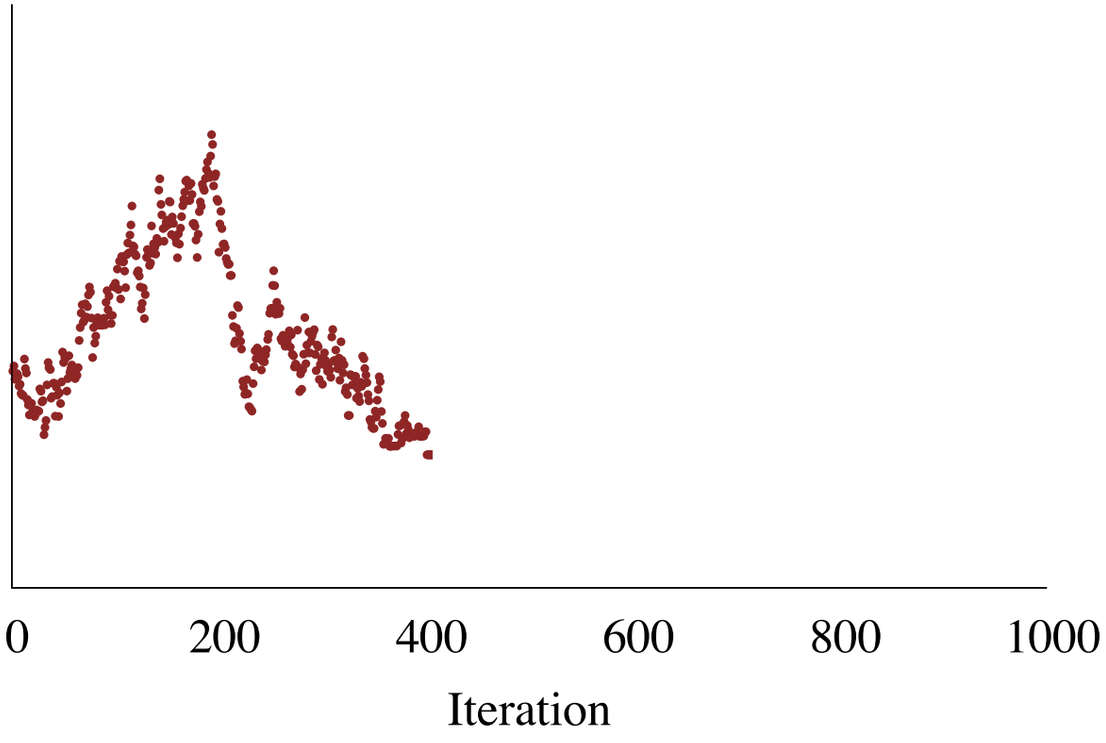


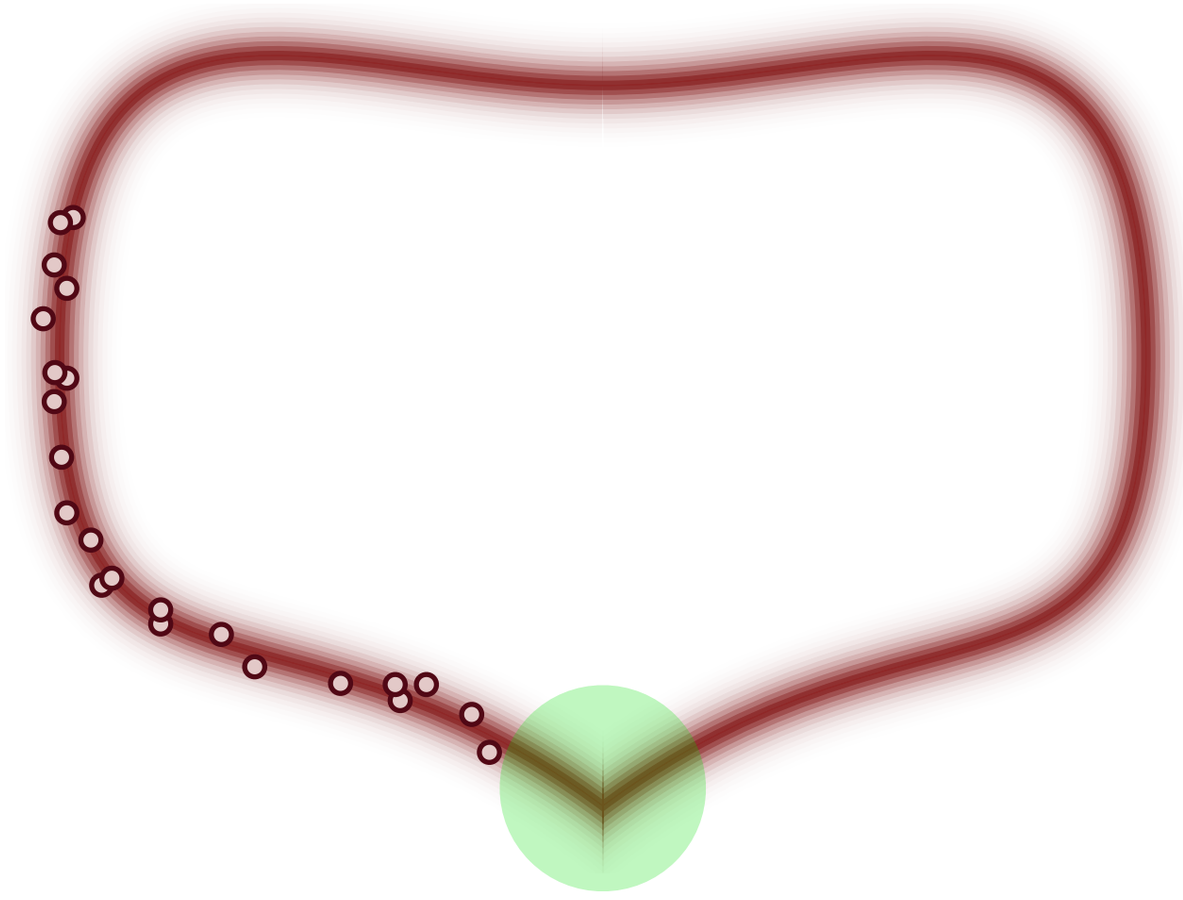
We can achieve a computationally efficient implementation, but how robust will it be?



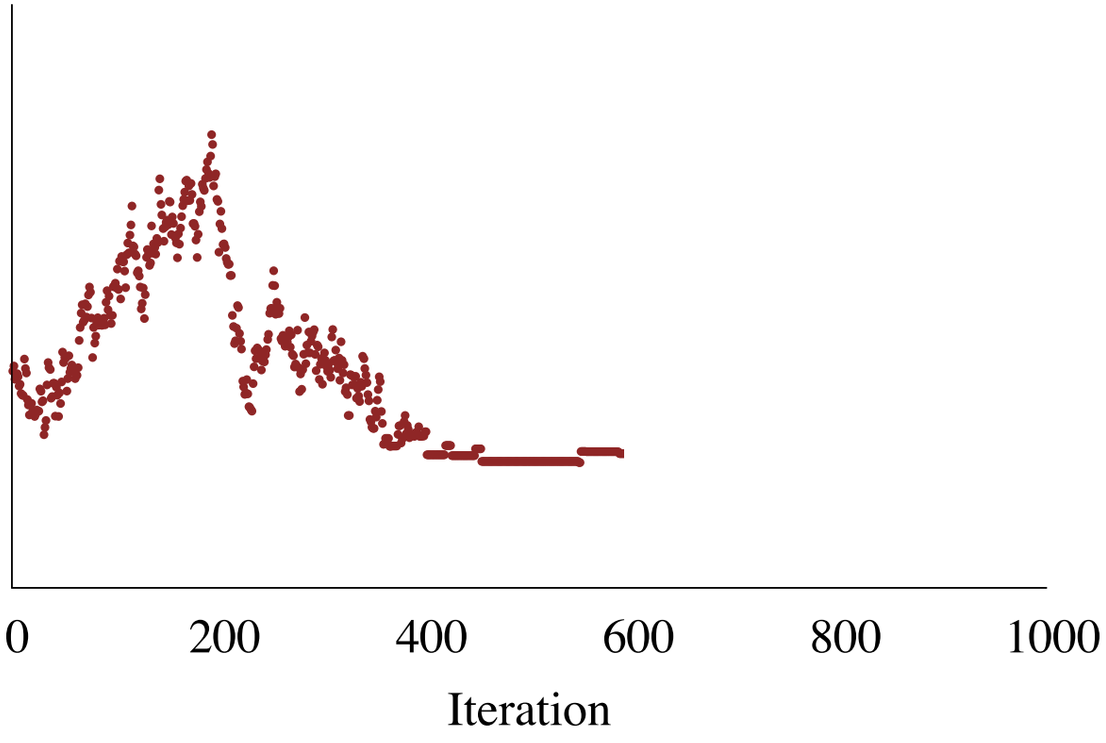


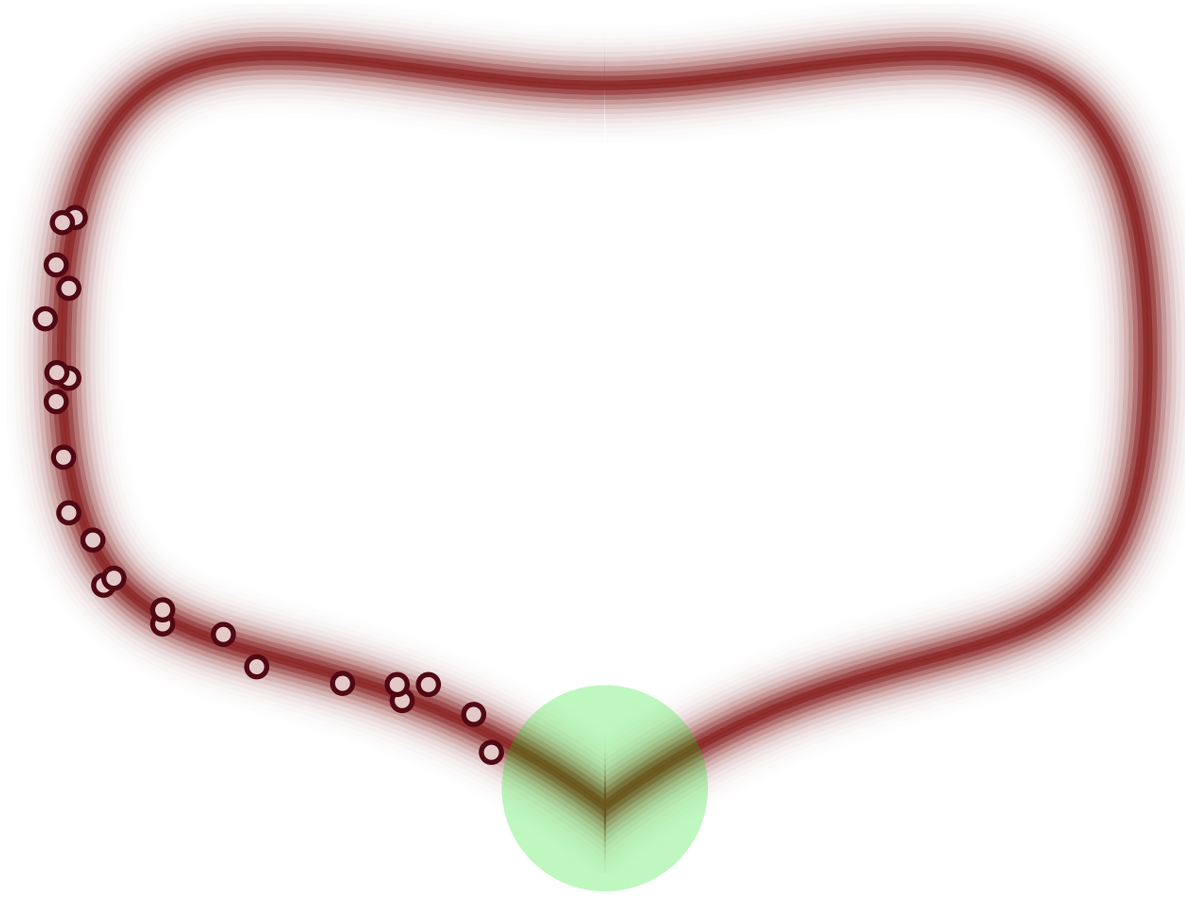
We can achieve a computationally efficient implementation, but how robust will it be?



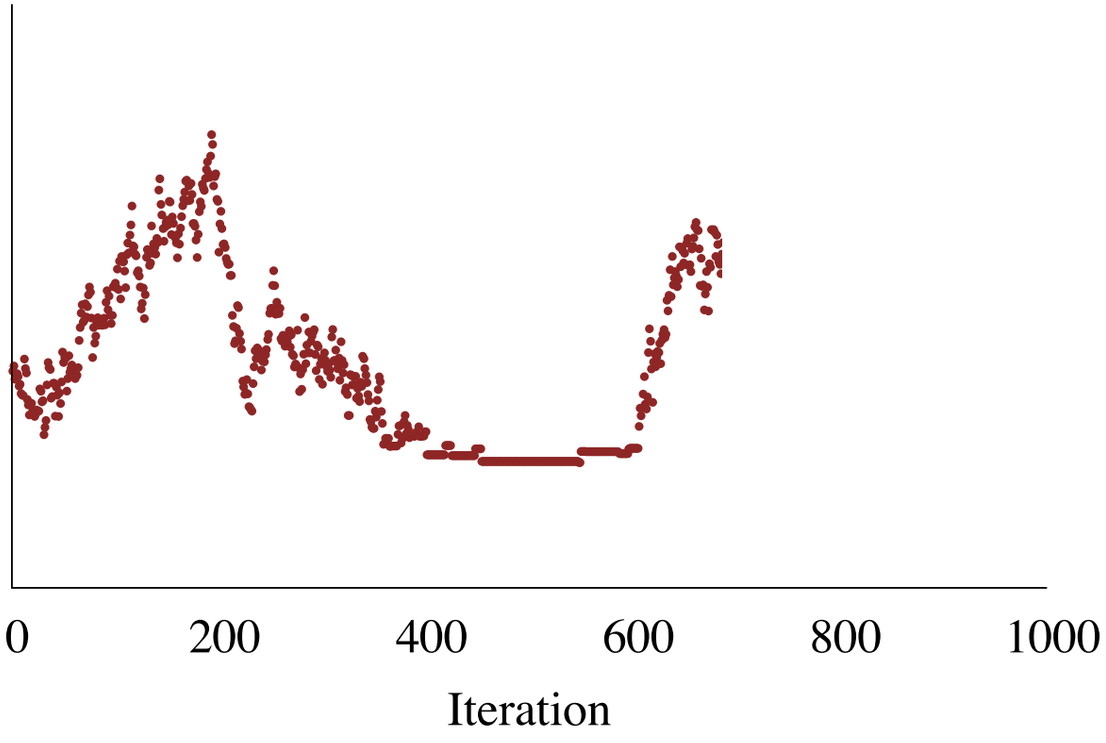


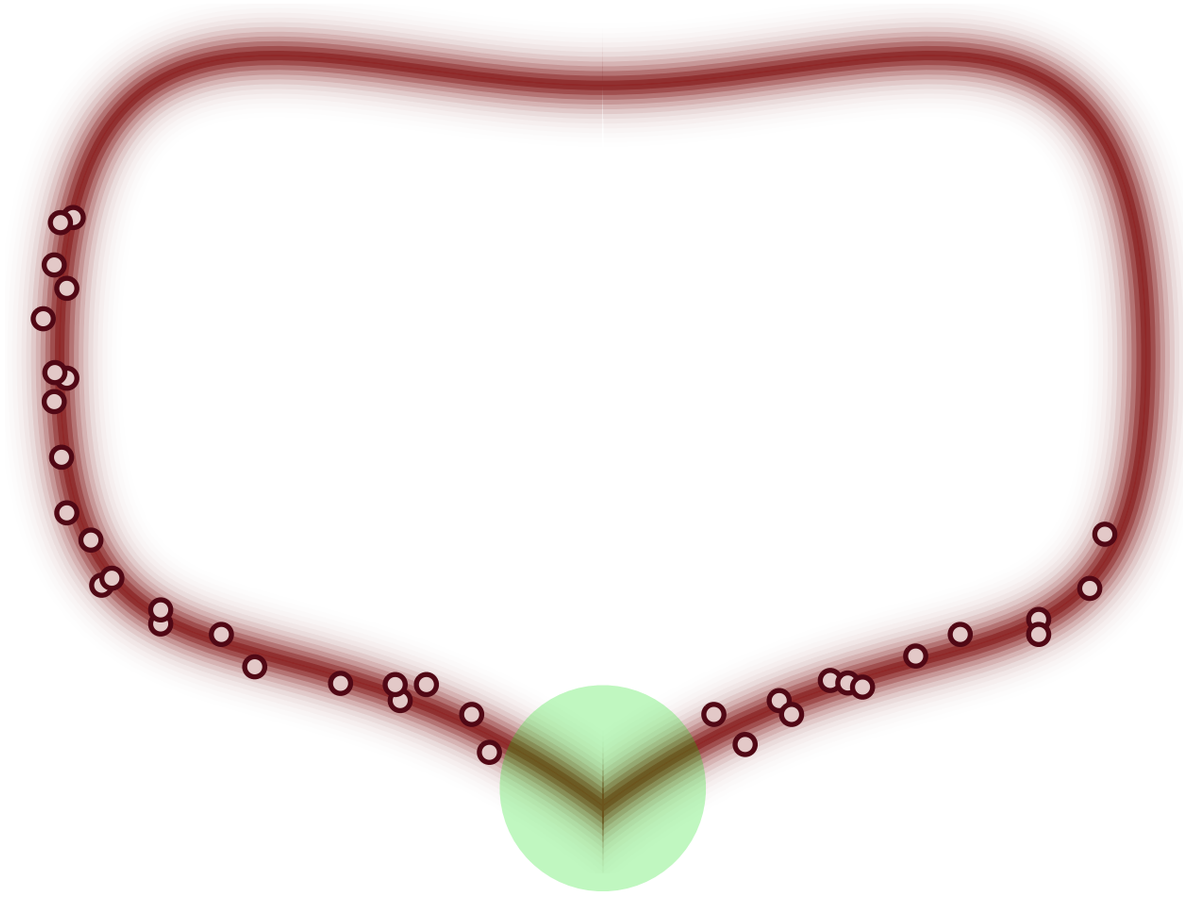
We can achieve a computationally efficient implementation, but how robust will it be?



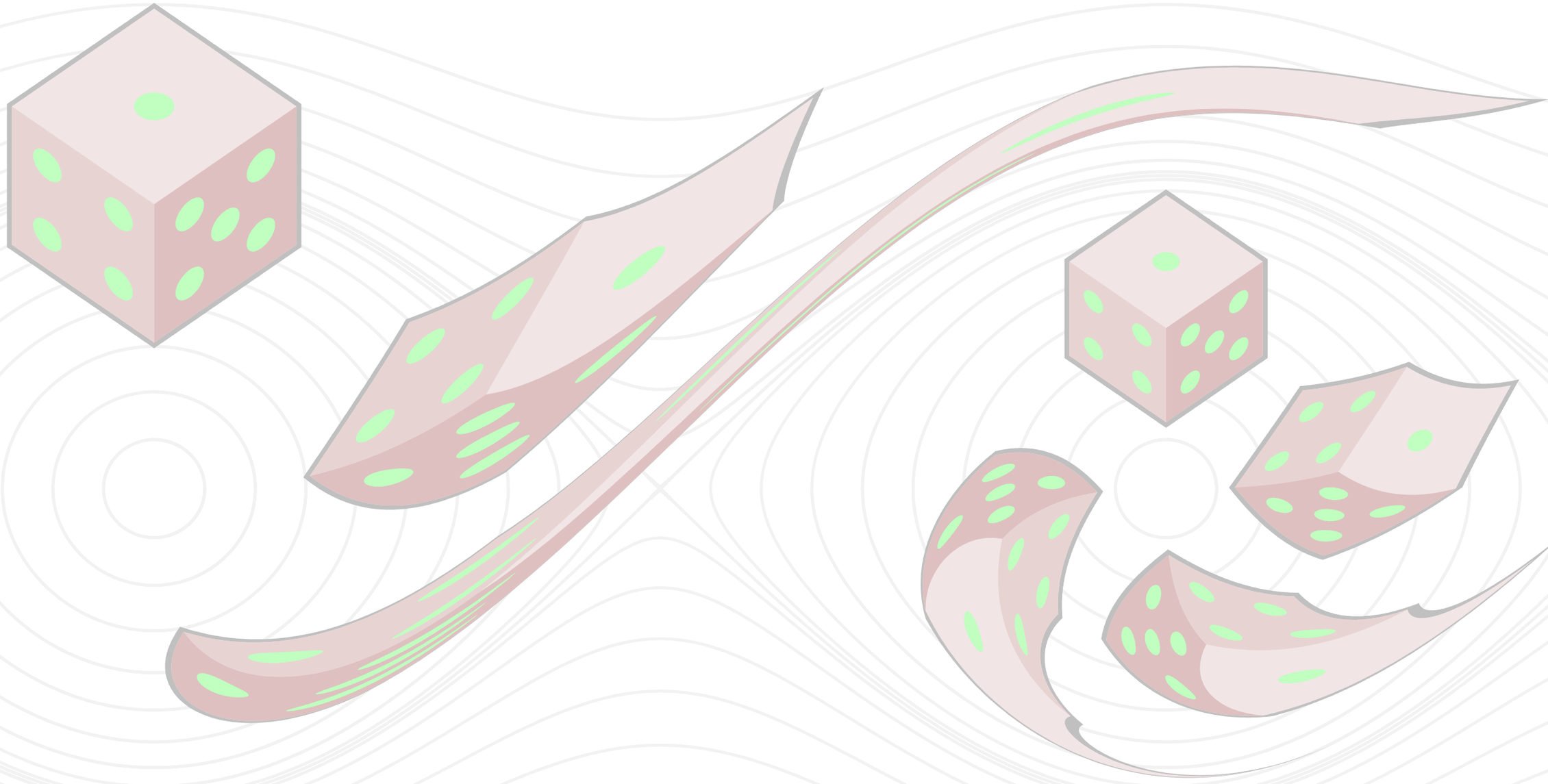


We can achieve a computationally efficient implementation, but how robust will it be?





The Theoretical Foundations of Hamiltonian Monte Carlo



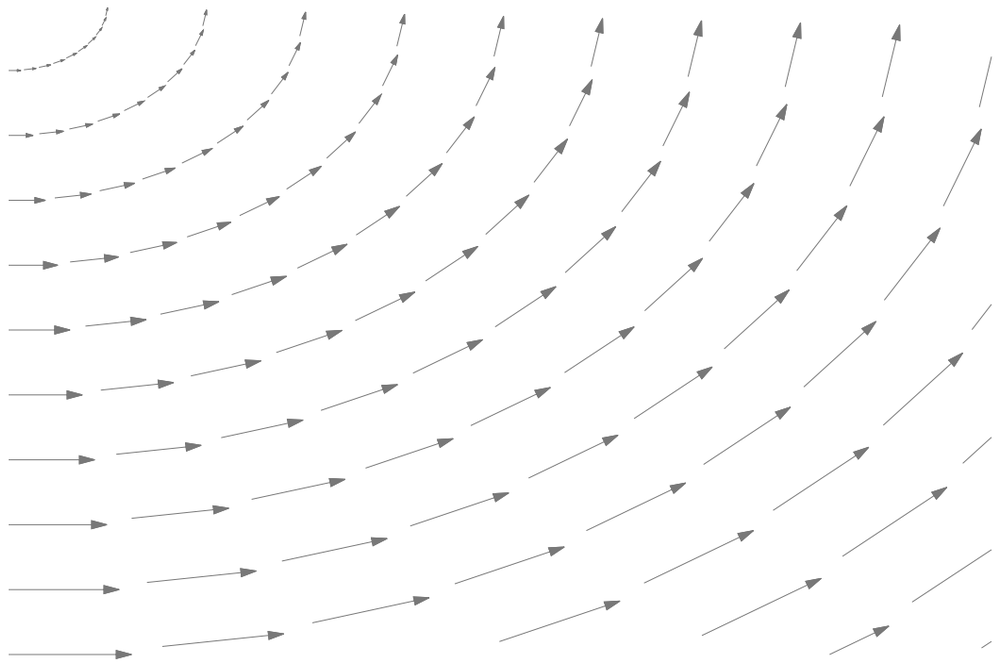
Measure-preserving flows arise naturally in *Hamiltonian systems*.

(M, ω, H)

Measure-preserving flows arise naturally in *Hamiltonian systems.*

$$(M, \omega, H)$$

$$\mathrm{d}H = \omega(\vec{X}_H, \cdot)$$

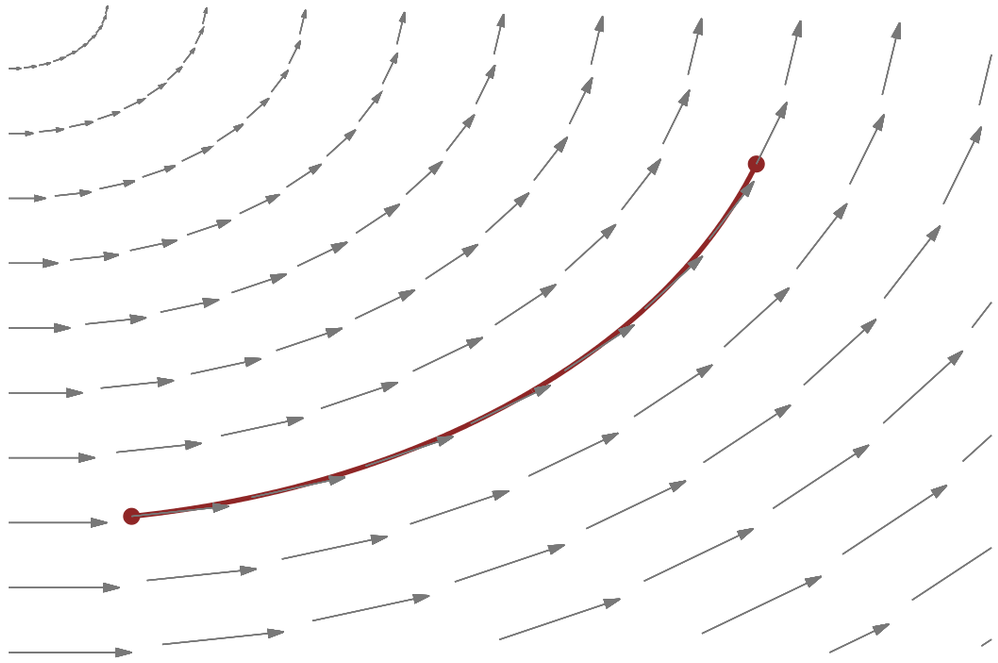


Measure-preserving flows arise naturally in *Hamiltonian systems*.

$$(M, \omega, H)$$

$$\mathrm{d}H = \omega(\vec{X}_H, \cdot)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t^H = \vec{X}_H$$

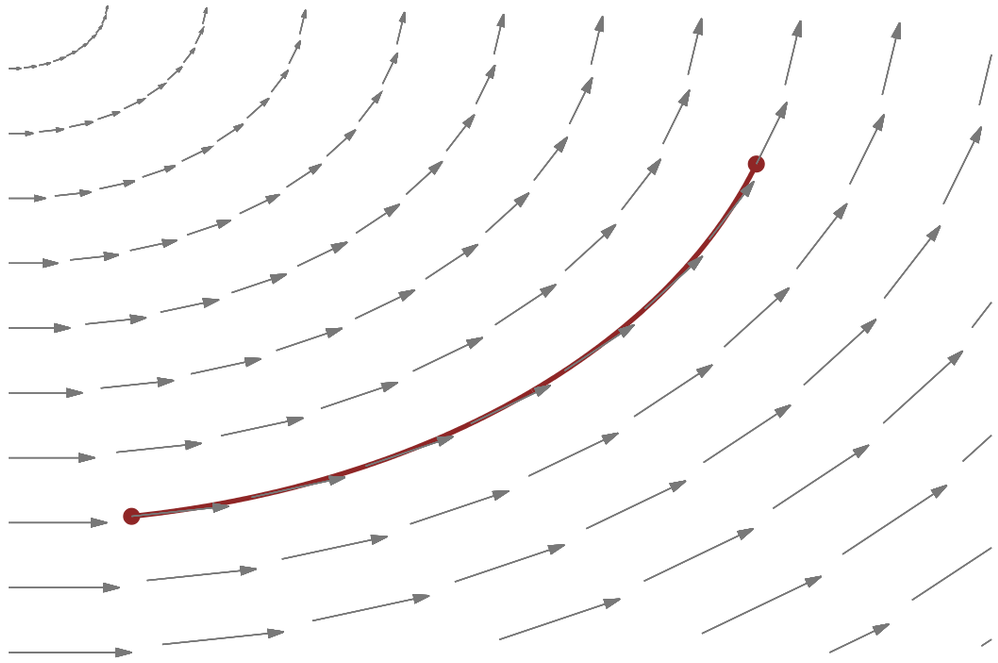


Measure-preserving flows arise naturally in *Hamiltonian systems*.

$$(M, \omega, H)$$

$$\mathrm{d}H = \omega(\vec{X}_H, \cdot)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t^H = \vec{X}_H$$



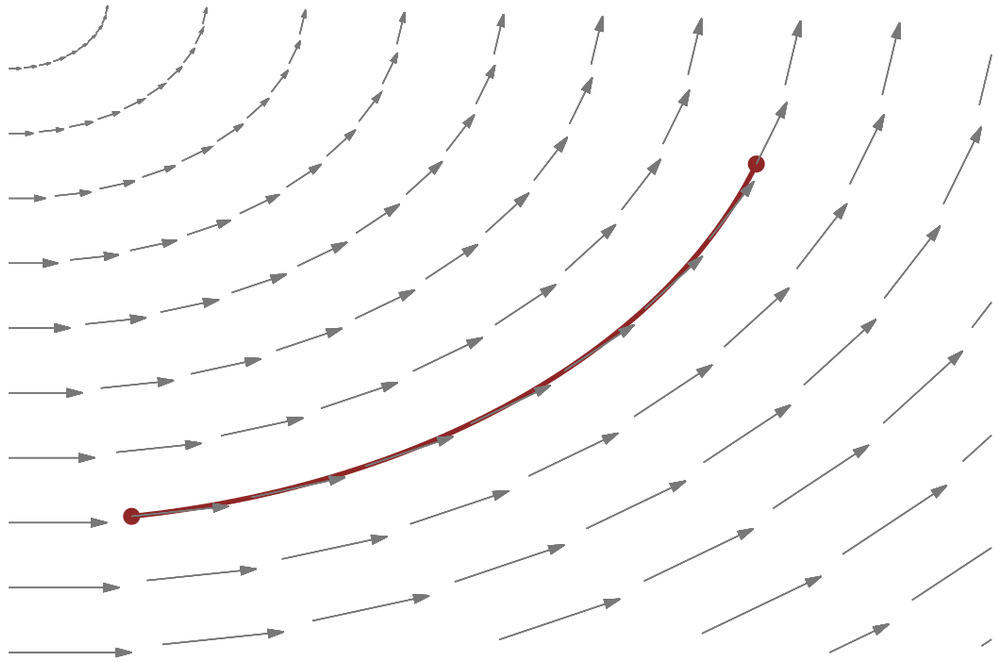
$$\omega^n$$

Measure-preserving flows arise naturally in *Hamiltonian systems*.

$$(M, \omega, H)$$

$$\mathrm{d}H = \omega(\vec{X}_H, \cdot)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t^H = \vec{X}_H$$



$$\pi_H \propto e^{-H}\omega^n$$

Measure-preserving flows arise naturally in *Hamiltonian systems*.

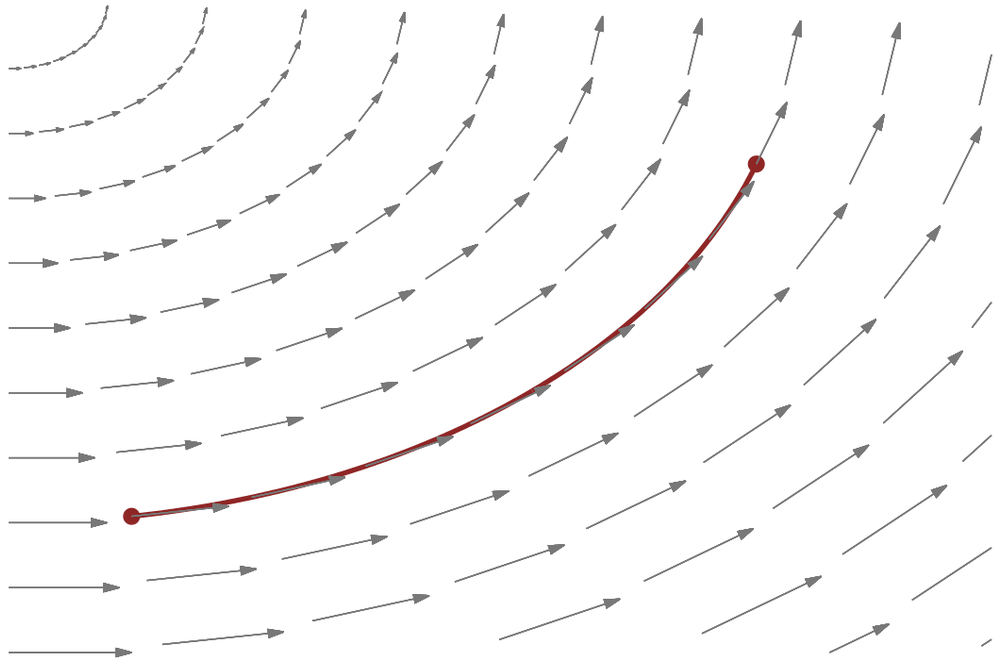
$$(M, \omega, H)$$

$$\mathrm{d}H = \omega(\vec{X}_H, \cdot)$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t^H = \vec{X}_H$$

$$\pi_H \propto e^{-H}\omega^n$$

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

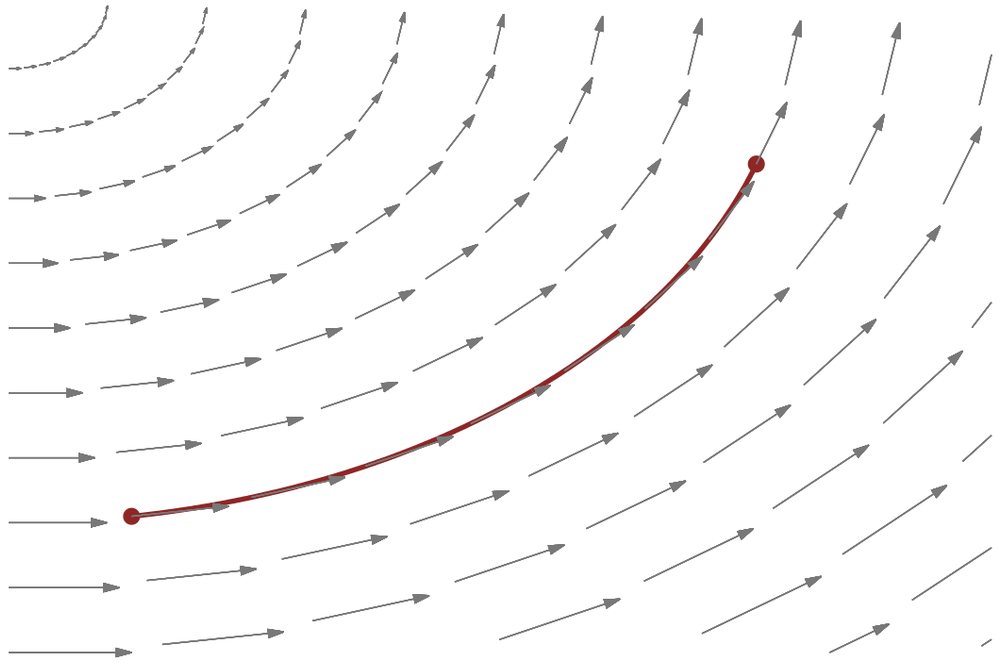


Unfortunately, an arbitrary target space doesn't have the necessary structure to construct a Hamiltonian system.

$$(M, \omega, H)$$

$$\mathrm{d}H = \omega(\vec{X}_H, \cdot)$$

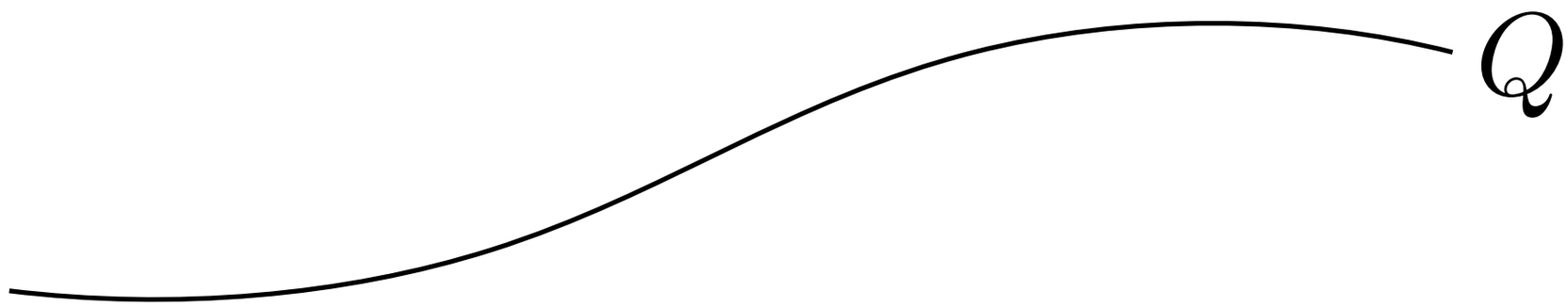
$$\frac{\mathrm{d}}{\mathrm{d}t}\phi_t^H = \vec{X}_H$$



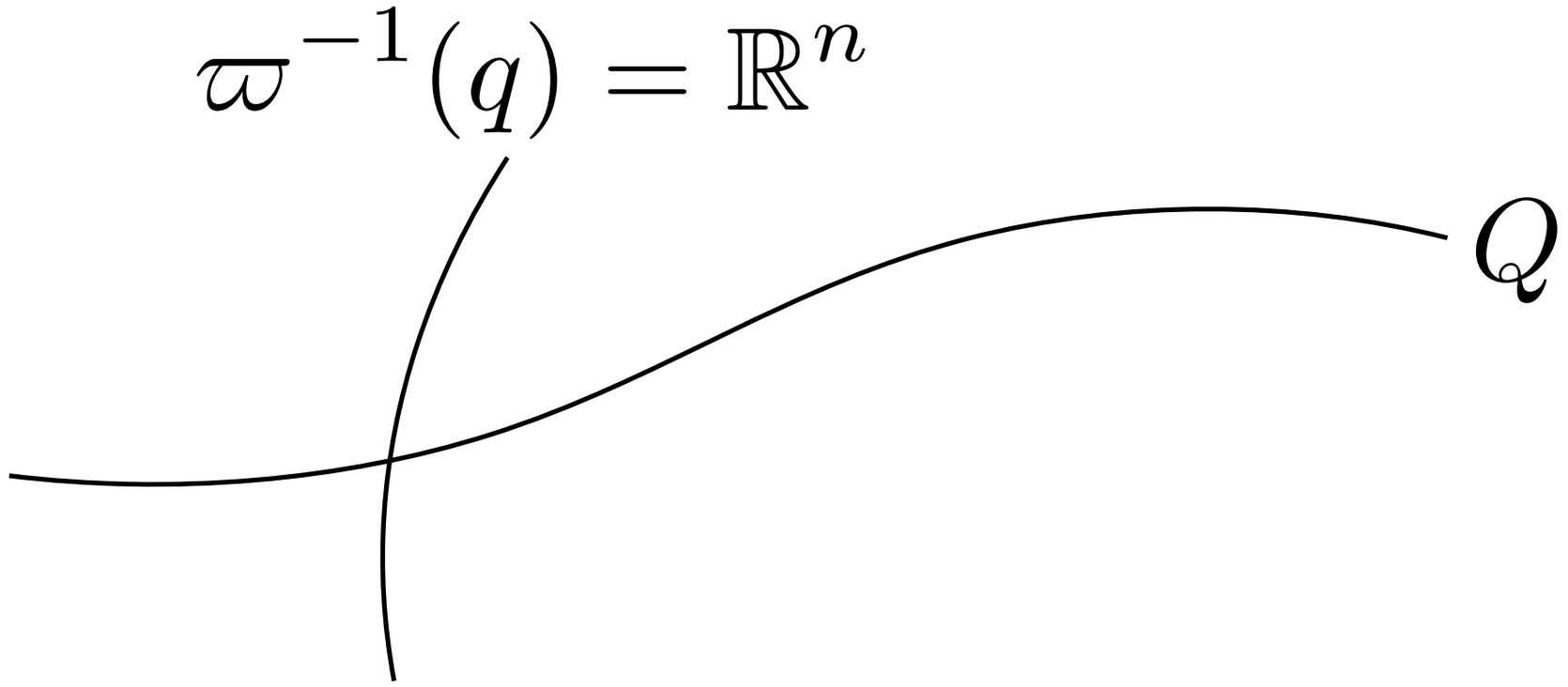
$$\pi_H \propto e^{-H}\omega^n$$

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

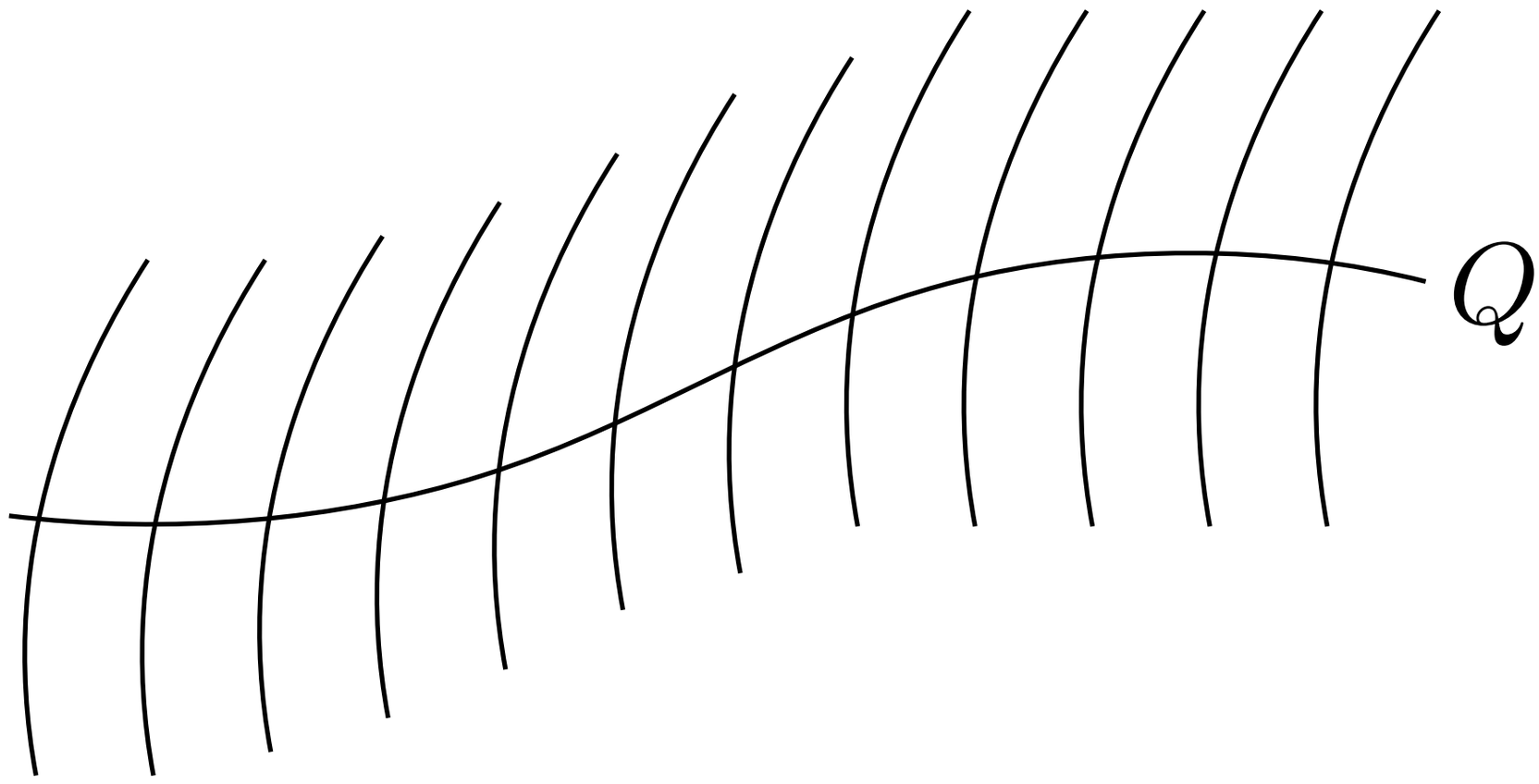
But the *cotangent bundle* of a smooth target space does!



But the *cotangent bundle* of a smooth target space does!

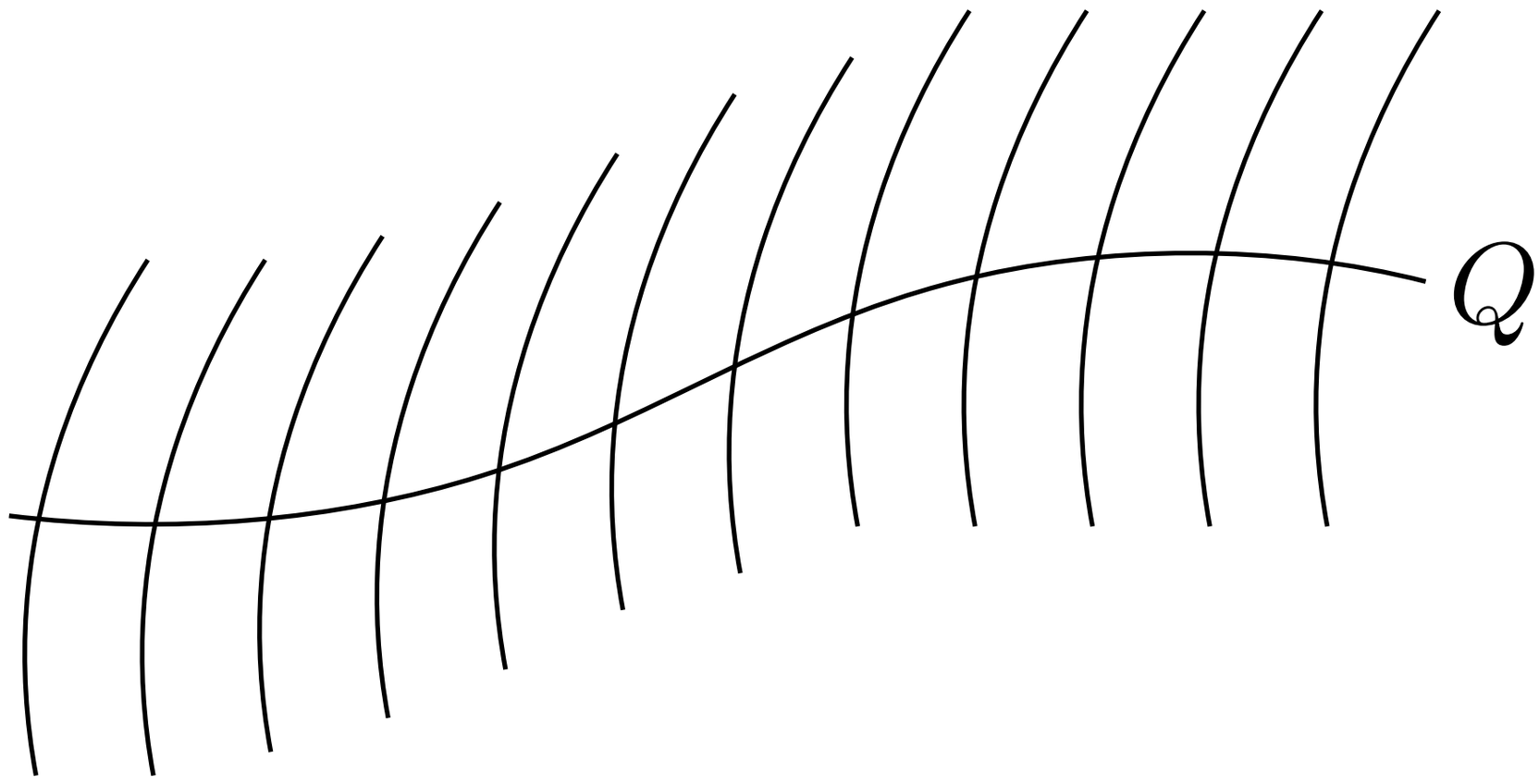


But the *cotangent bundle* of a smooth target space does!



$$\varpi : T^*Q \to Q$$

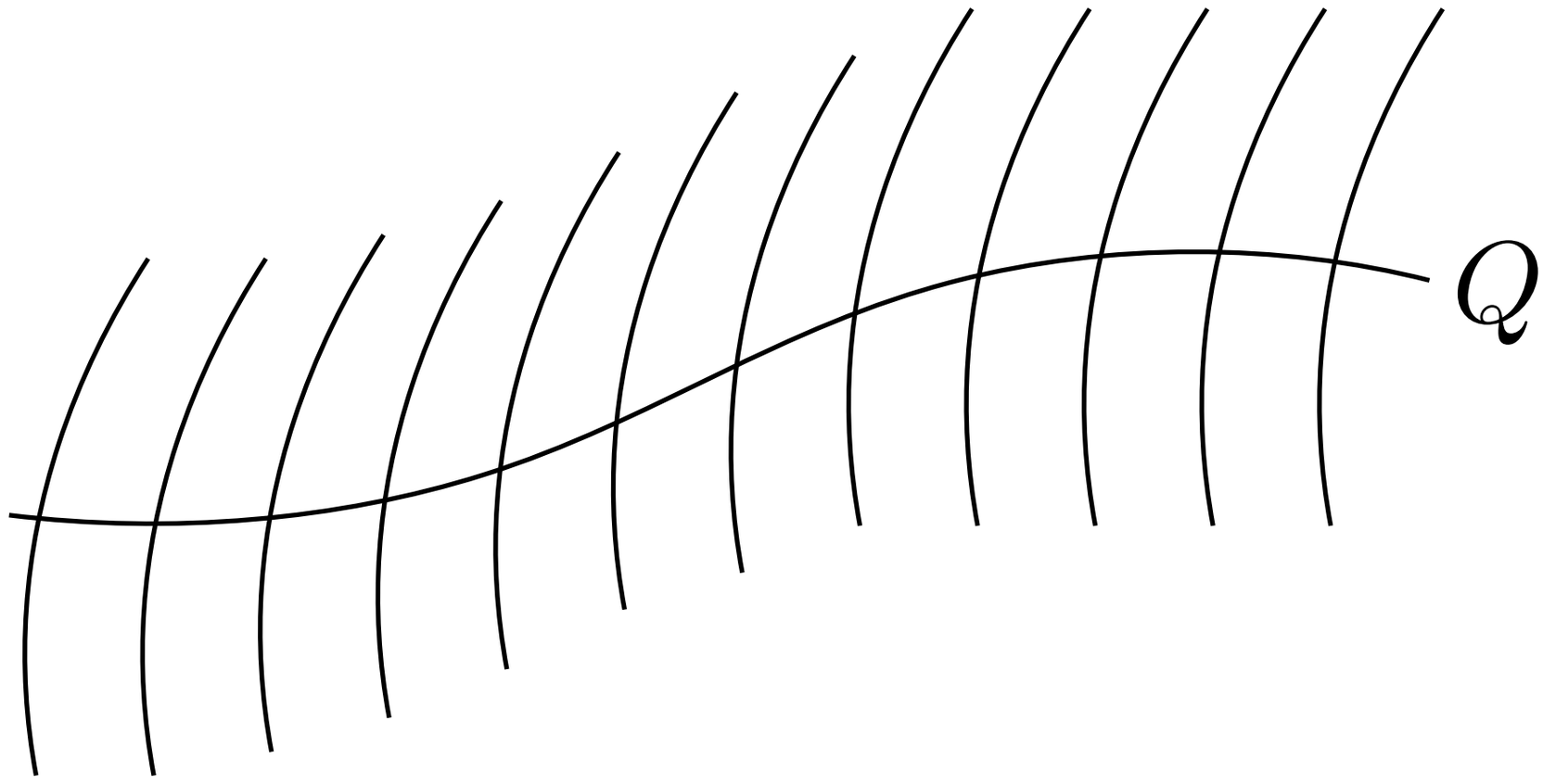
But the *cotangent bundle* of a smooth target space does!



$$\varpi : T^*Q \to Q$$

$$(T^*Q, \omega)$$

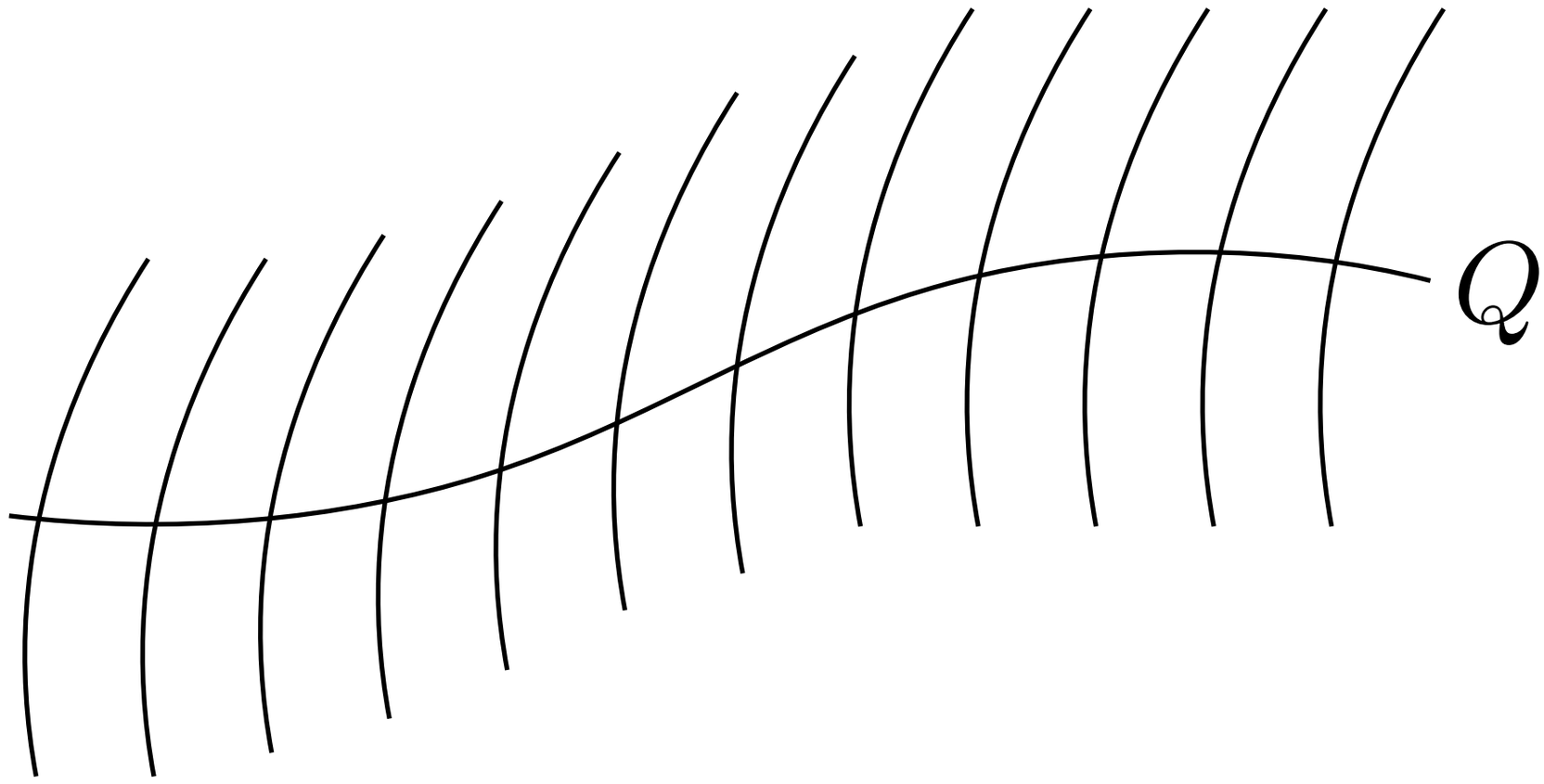
But the *cotangent bundle* of a smooth target space does!



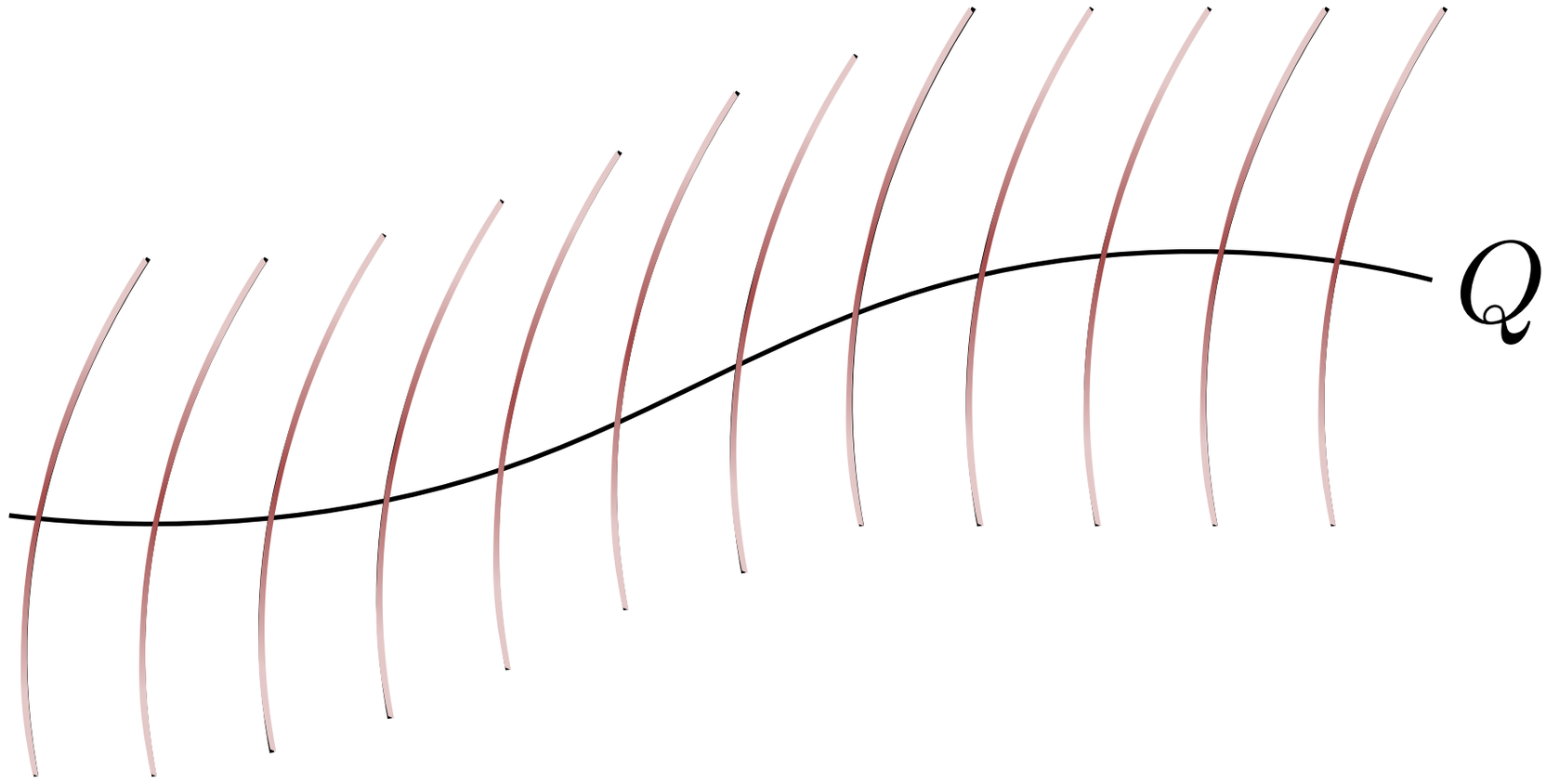
$$\varpi : T^*Q \to Q$$

$$(T^*Q, \omega, H?)$$

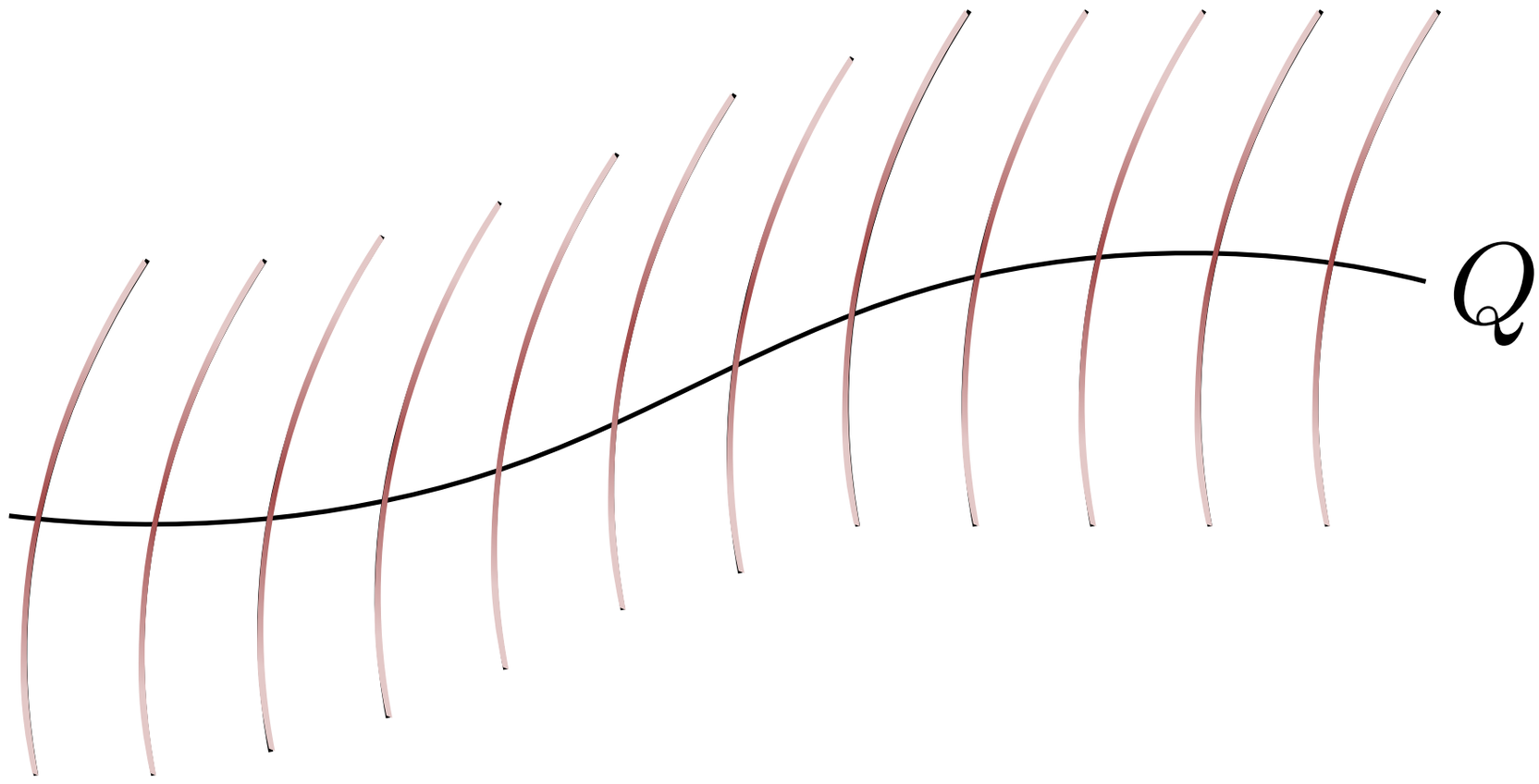
In order to map the given probabilistic system into a Hamiltonian one we need to select a *disintegration*.



In order to map the given probabilistic system into a Hamiltonian one we need to select a *disintegration*.



In order to map the given probabilistic system into a Hamiltonian one we need to select a *disintegration*.



$$\xi = \pi(\mathrm{d}p|q) = e^{-T(p,q)}\mathrm{d}p$$

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

Q

$\varpi : T^*Q \to Q$

π

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

Q

$\varpi : T^*Q \to Q$

$\pi \longrightarrow \pi_H = \varpi^* \pi \wedge \xi$

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

Q

$\varpi : T^*Q \to Q$

$$\pi \longrightarrow \pi_H = \varpi^* \pi \wedge \xi$$

$$\downarrow$$

$$H = -\log \frac{\mathrm{d}\pi_H}{\mathrm{d}\omega^n}$$

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

Q

$\varpi : T^*Q \to Q$

π $\longrightarrow$ $\pi_H = \varpi^* \pi \wedge \xi$

$\downarrow$

$H = -\log \dfrac{\mathrm{d}\pi_H}{\mathrm{d}\omega^n}$

$\downarrow$

ϕ_t^H

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

Q

$\varpi : T^*Q \to Q$

π $\longrightarrow$ $\pi_H = \varpi^*\pi \wedge \xi$

$\downarrow$

$H = -\log \dfrac{\mathrm{d}\pi_H}{\mathrm{d}\omega^n}$

$\downarrow$

$\varpi \circ \phi_t^H$ $\longleftarrow$ ϕ_t^H

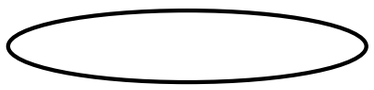
Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

$$\frac{dq}{dt} = \frac{\partial T}{\partial p}$$

$$\frac{dp}{dt} = -\frac{\partial T}{\partial q} - \frac{\partial V}{\partial q}$$

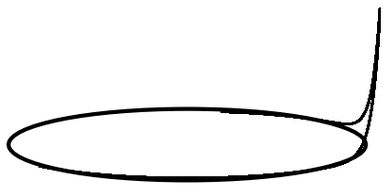
$$\phi_t^H$$

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



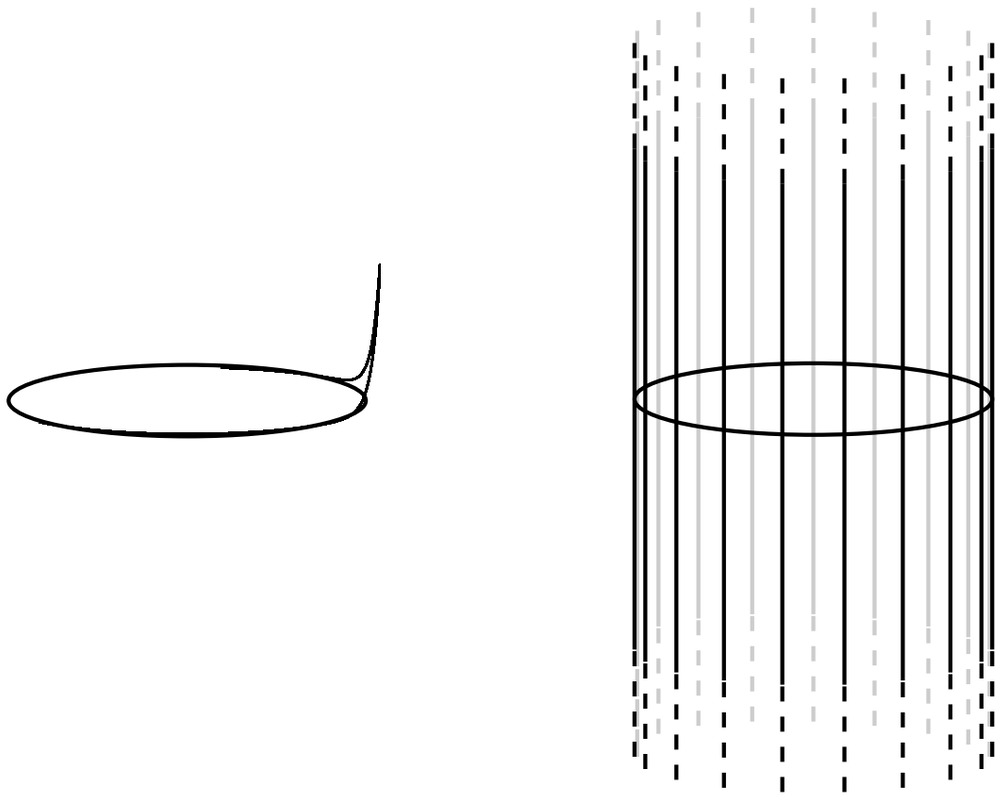
$Q = \mathbb{S}^1$

Hamiltonian Monte Carlo uses the cotangent bundle
to construct the desired measure-preserving flow.



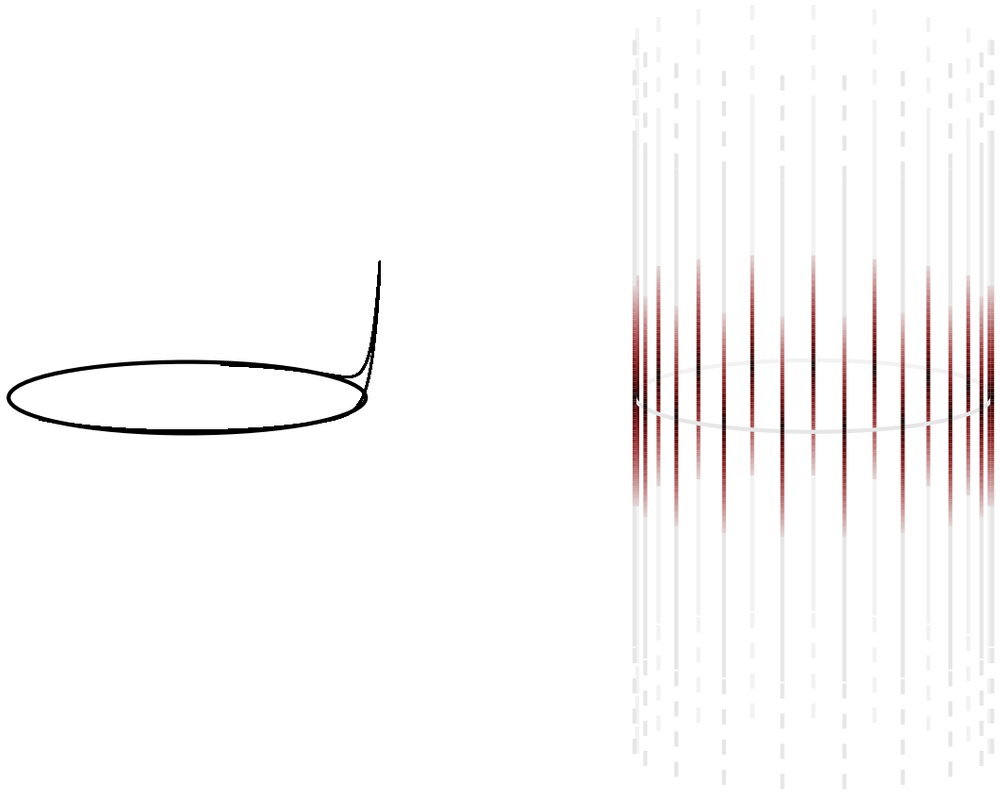
$$Q = \mathbb{S}^1$$

Hamiltonian Monte Carlo uses the cotangent bundle
to construct the desired measure-preserving flow.



$Q = \mathbb{S}^1$ $T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$

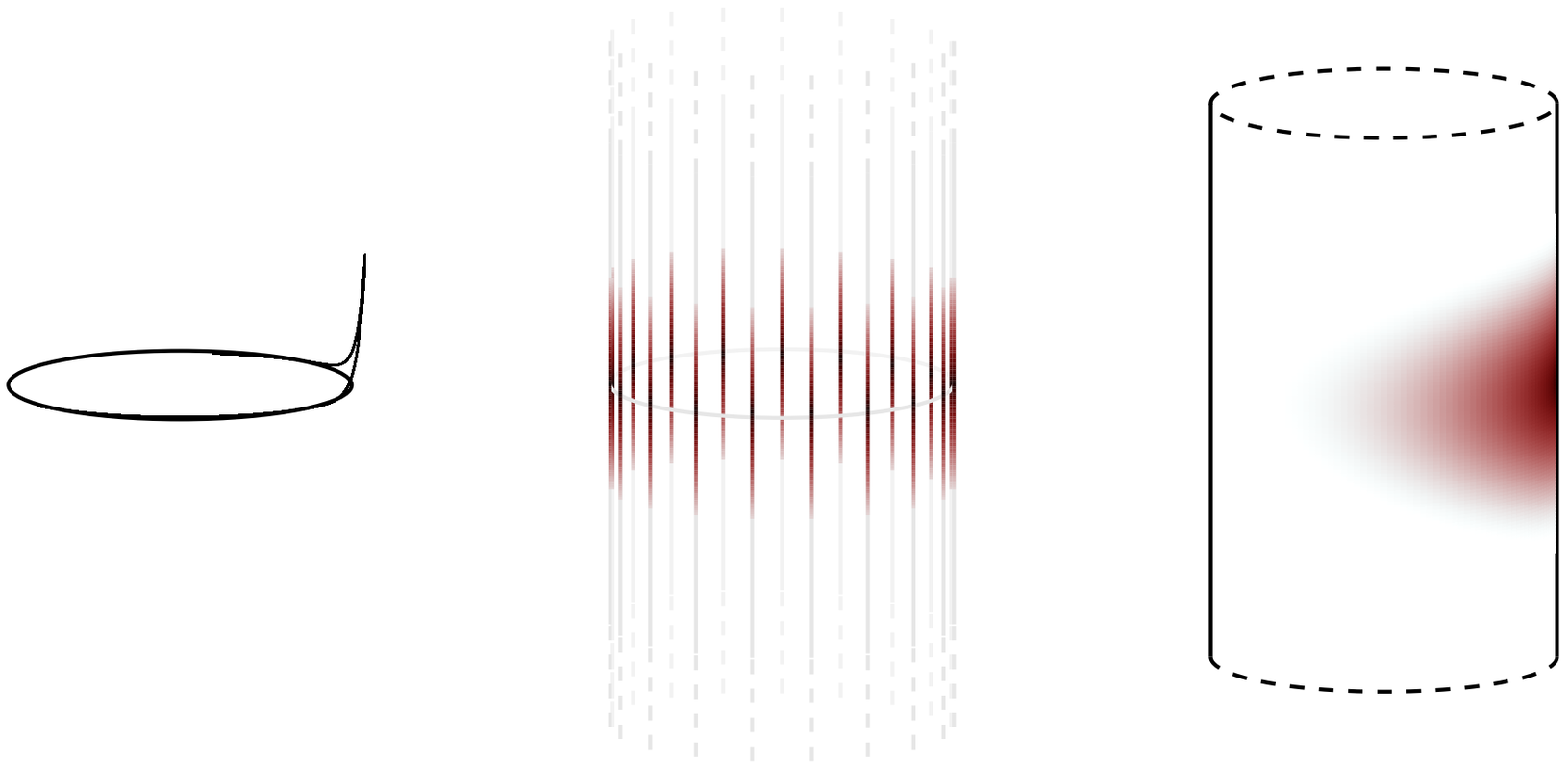
Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



$Q = \mathbb{S}^1$

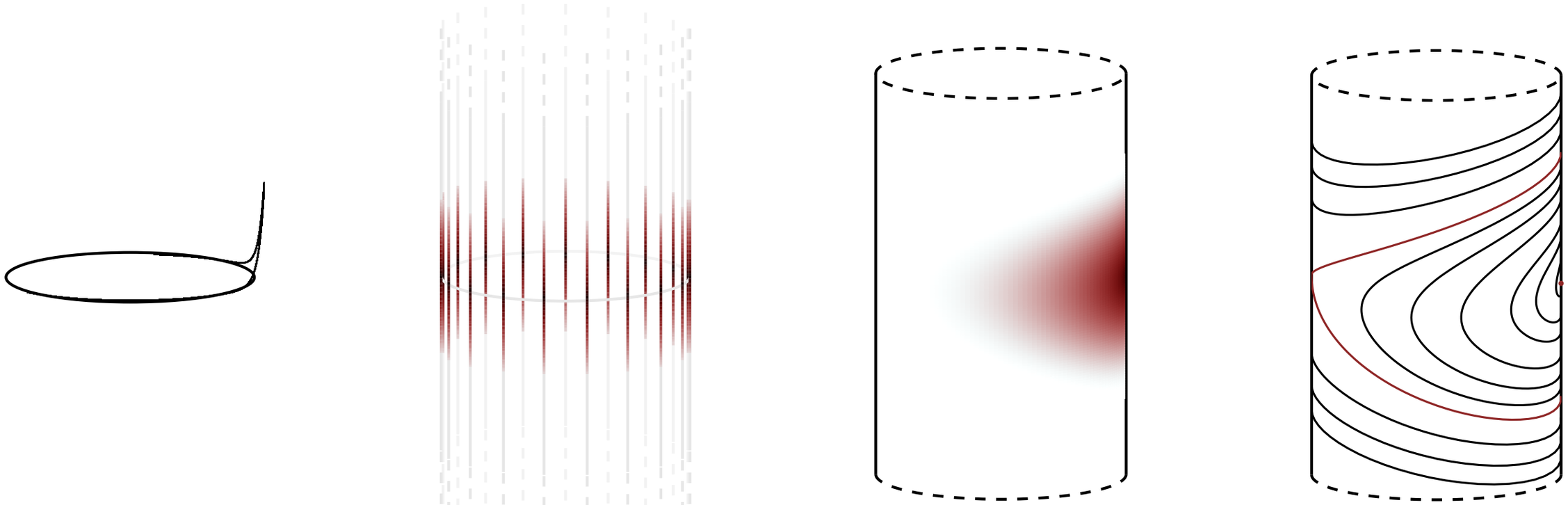
$T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



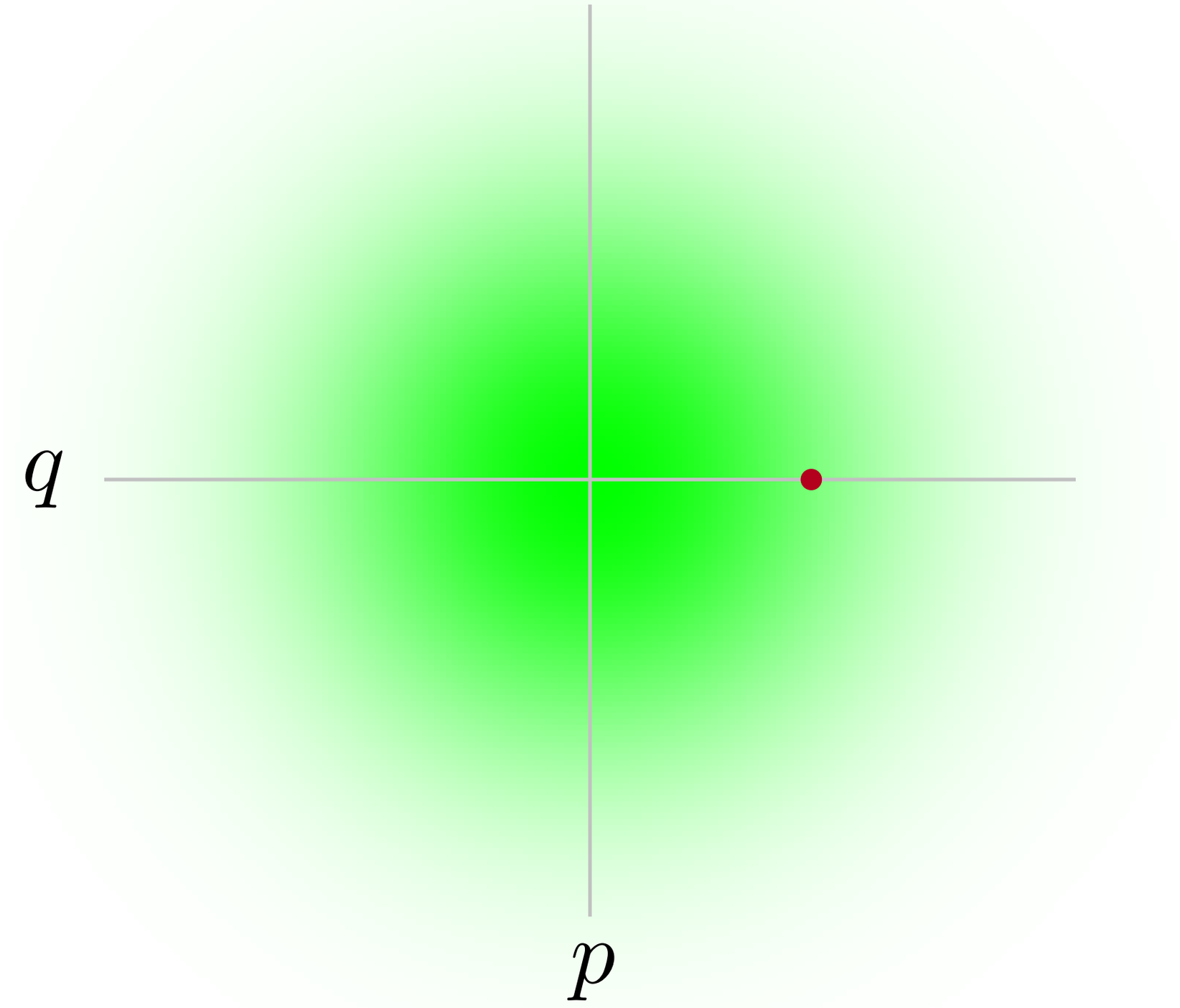
$Q = \mathbb{S}^1$ $T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$ $T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$

Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.

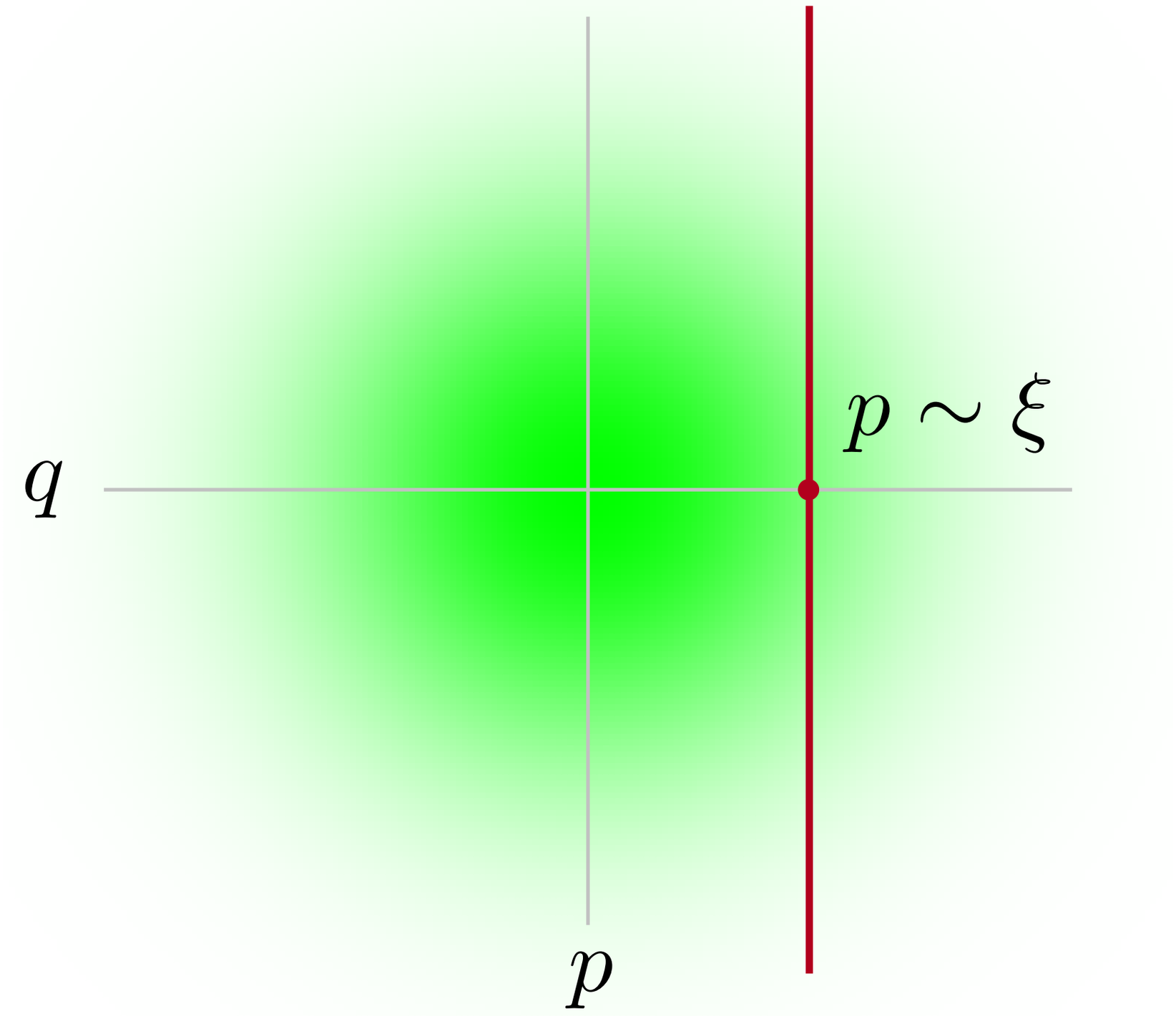


$Q = \mathbb{S}^1$ $T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$ $T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$ $T^*Q \approx \mathbb{S}^1 \times \mathbb{R}$

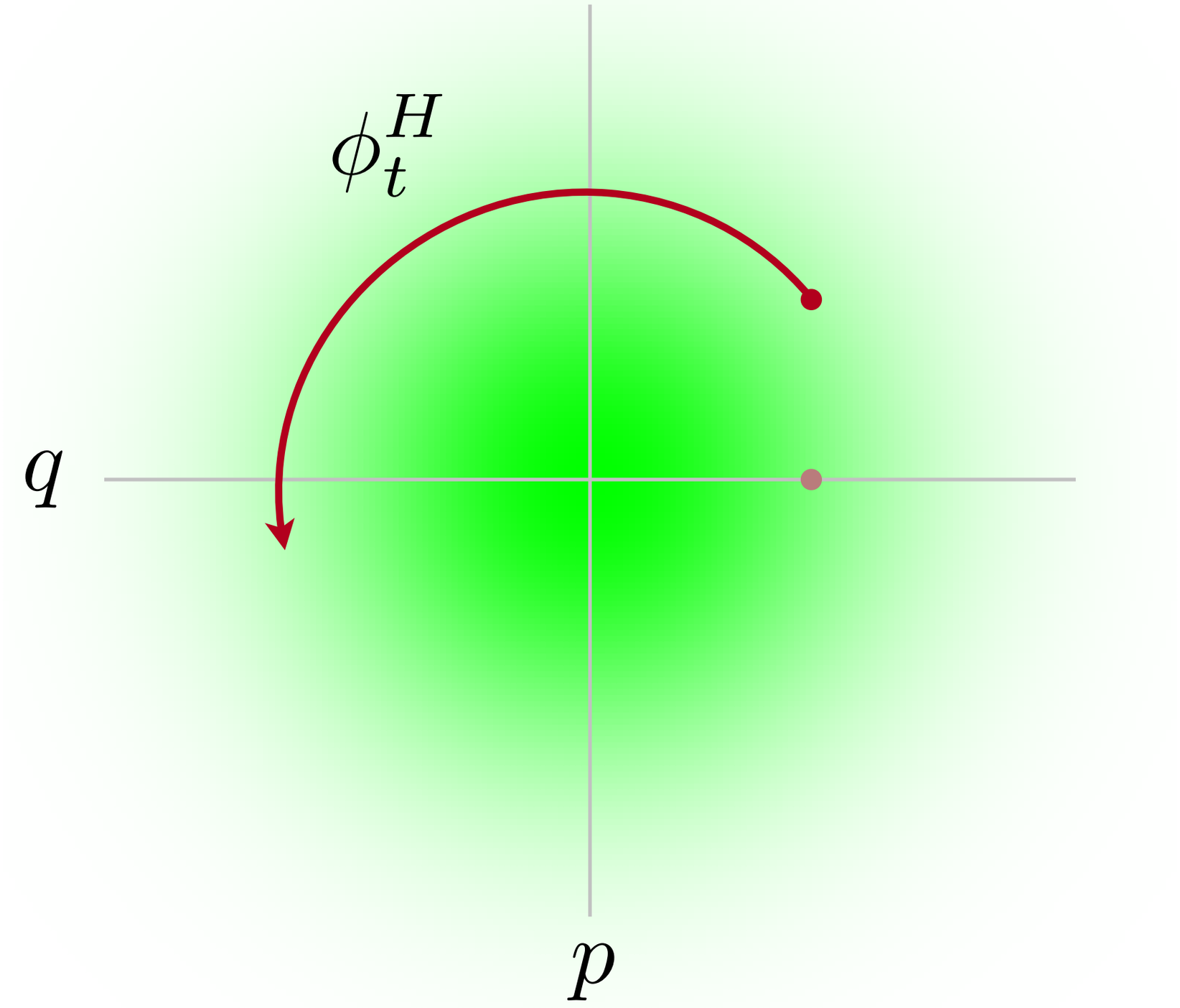
Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



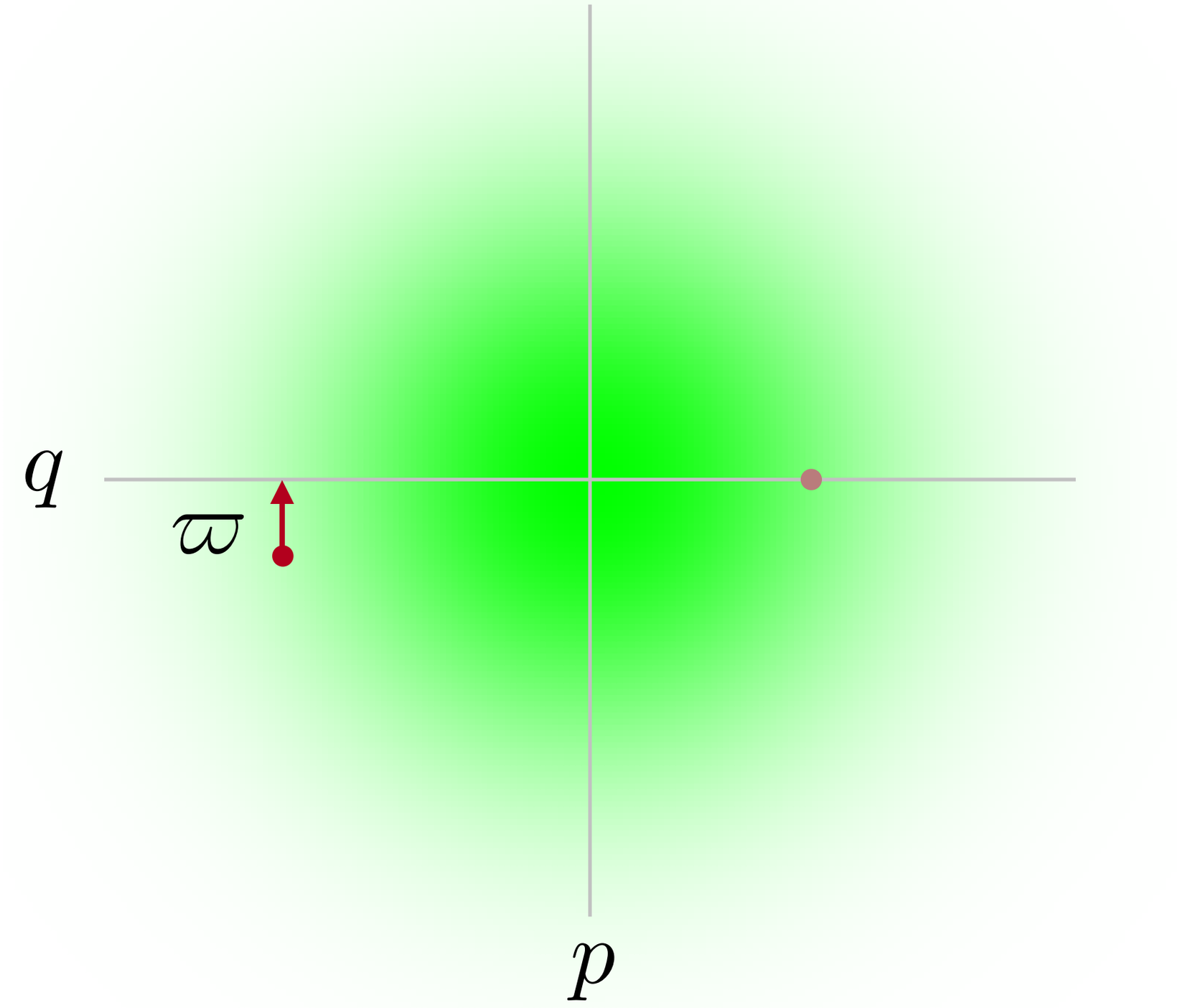
Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



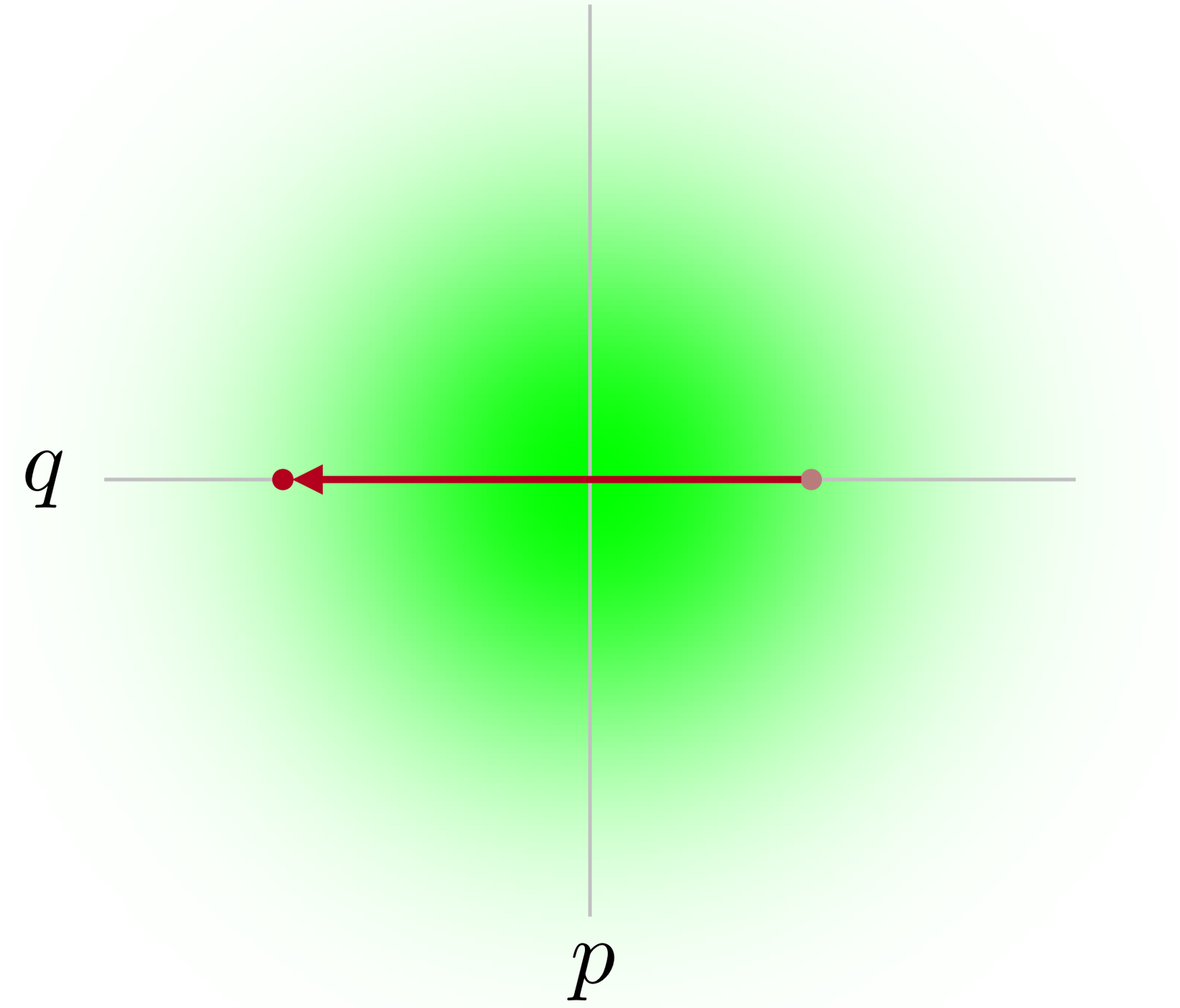
Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



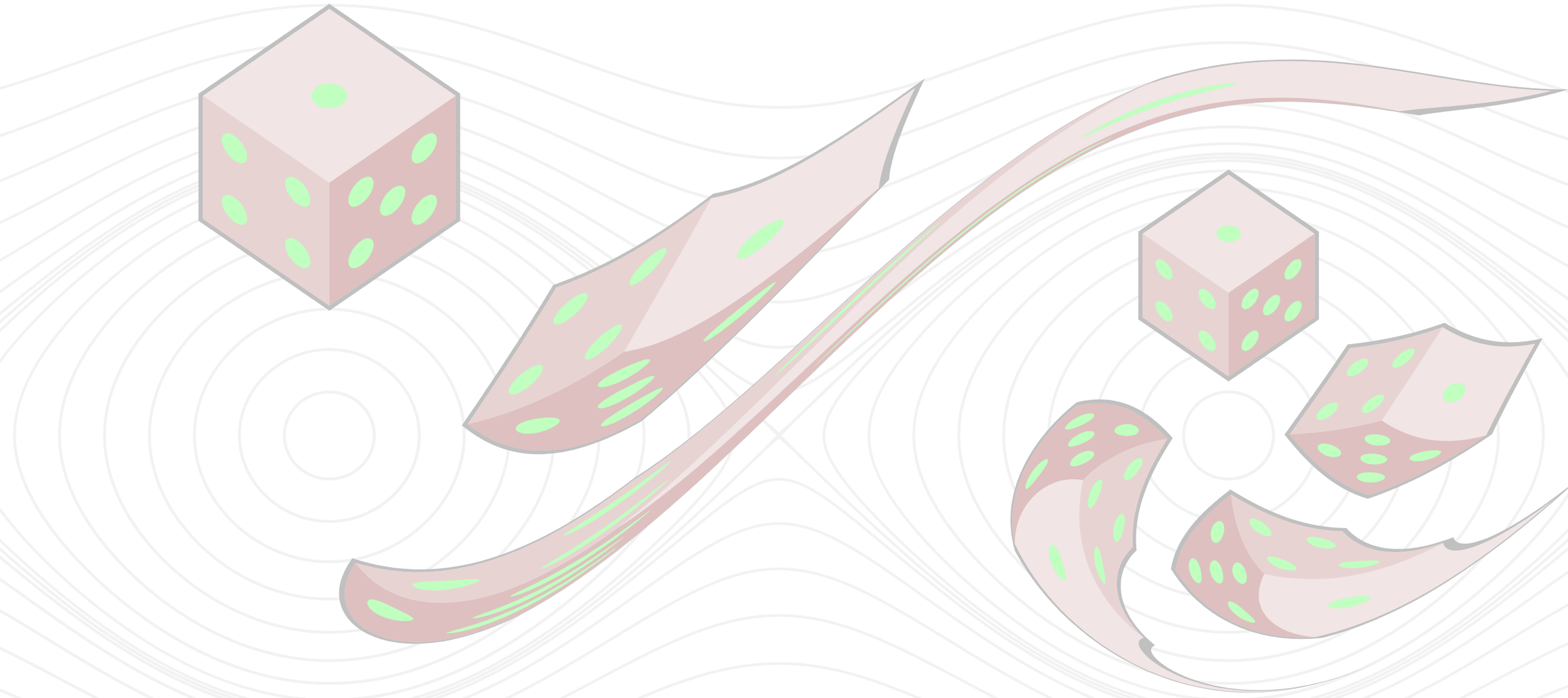
Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



Hamiltonian Monte Carlo uses the cotangent bundle to construct the desired measure-preserving flow.



The Dangers of Data Subsampling



Data subsampling is a popular strategy, but in complex models we need *all* of the data to explore the typical set.

Data subsampling is a popular strategy, but in complex models we need *all* of the data to explore the typical set.

Data subsampling is a popular strategy, but in complex models we need *all* of the data to explore the typical set.

Data subsampling compromises exploration and biases inferences unless the data are highly redundant.

Using the geometry of symplectic integrators we can derive formal expressions for this data subsampling bias.

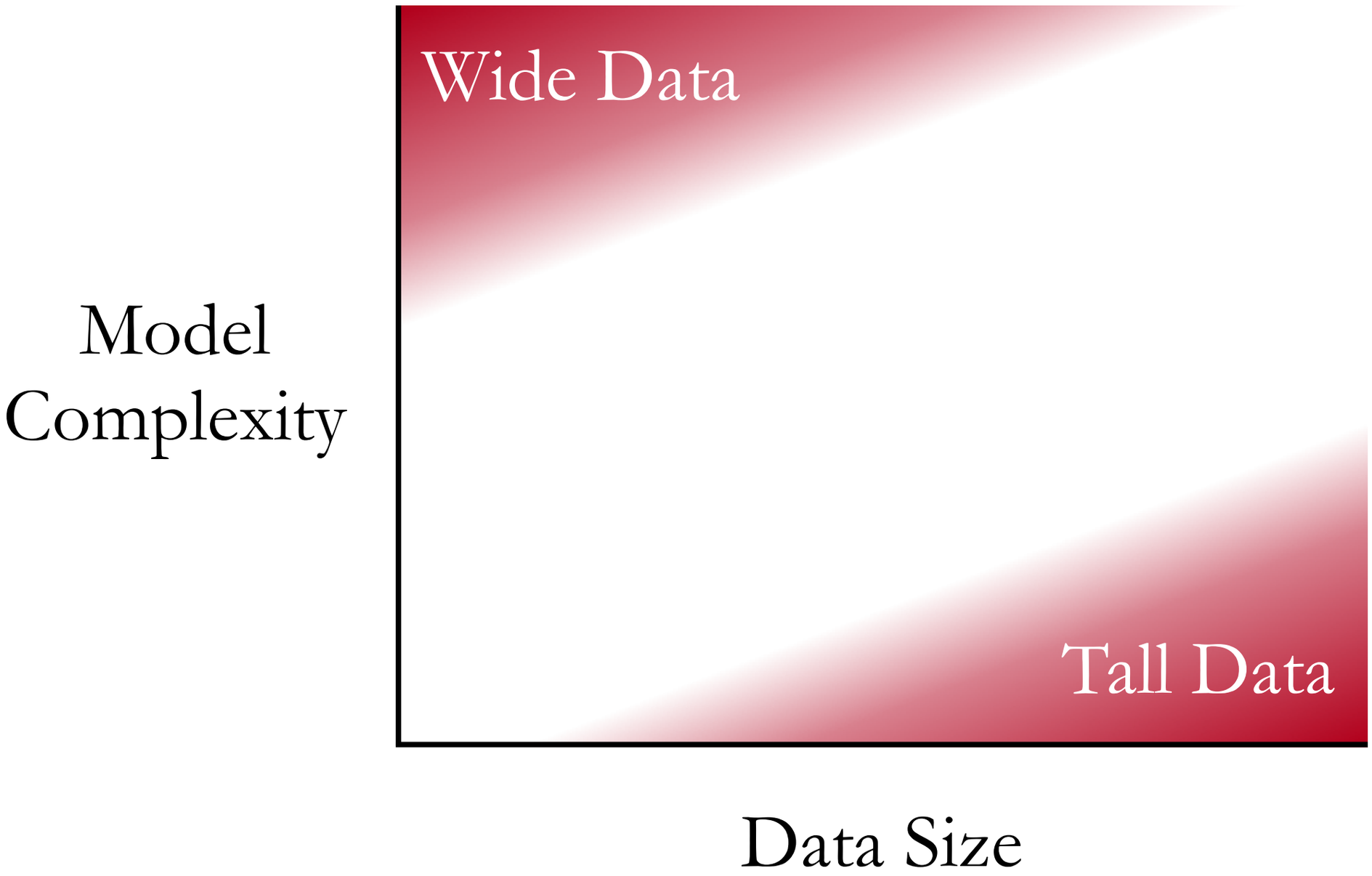
$$\left(e^{\frac{\epsilon}{2}J\vec{V}_j} \circ e^{\epsilon\vec{T}} \circ e^{\frac{\epsilon}{2}J\vec{V}_j}\right)^{\tau/\epsilon} = \exp\left(\tau\vec{H} - \tau\overrightarrow{\Delta V}_j\right) + \mathcal{O}(\epsilon^2)$$

Using the geometry of symplectic integrators we can derive formal expressions for this data subsampling bias.

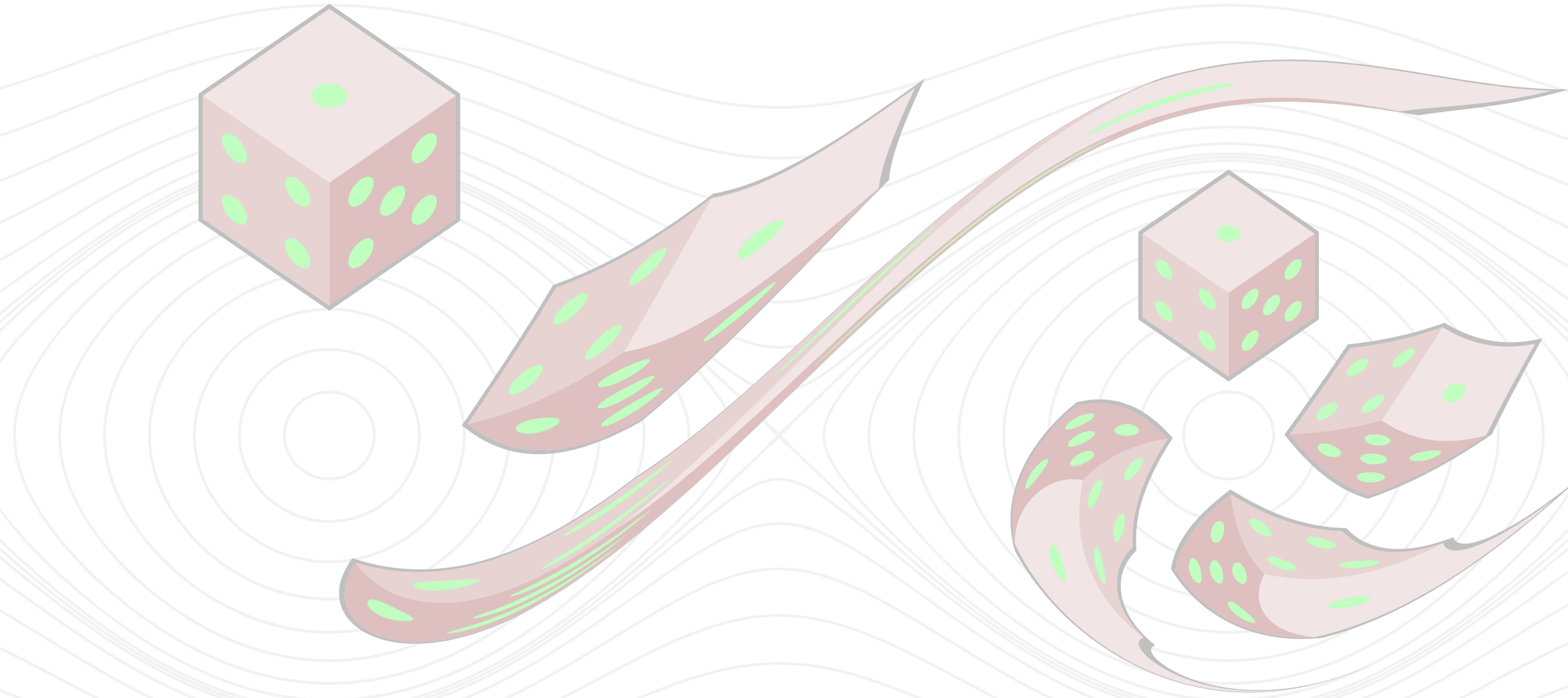
$$\left(e^{\frac{\epsilon}{2}J\vec{V}_j} \circ e^{\epsilon\vec{T}} \circ e^{\frac{\epsilon}{2}J\vec{V}_j}\right)^{\tau/\epsilon} = \exp\left(\tau\vec{H} - \tau\overrightarrow{\Delta V}_j\right) + \mathcal{O}(\epsilon^2)$$

$$\overrightarrow{\Delta V_j} = -\left(\frac{\partial V}{\partial q} - J\frac{\partial V_j}{\partial q}\right)\frac{\partial}{\partial p}$$

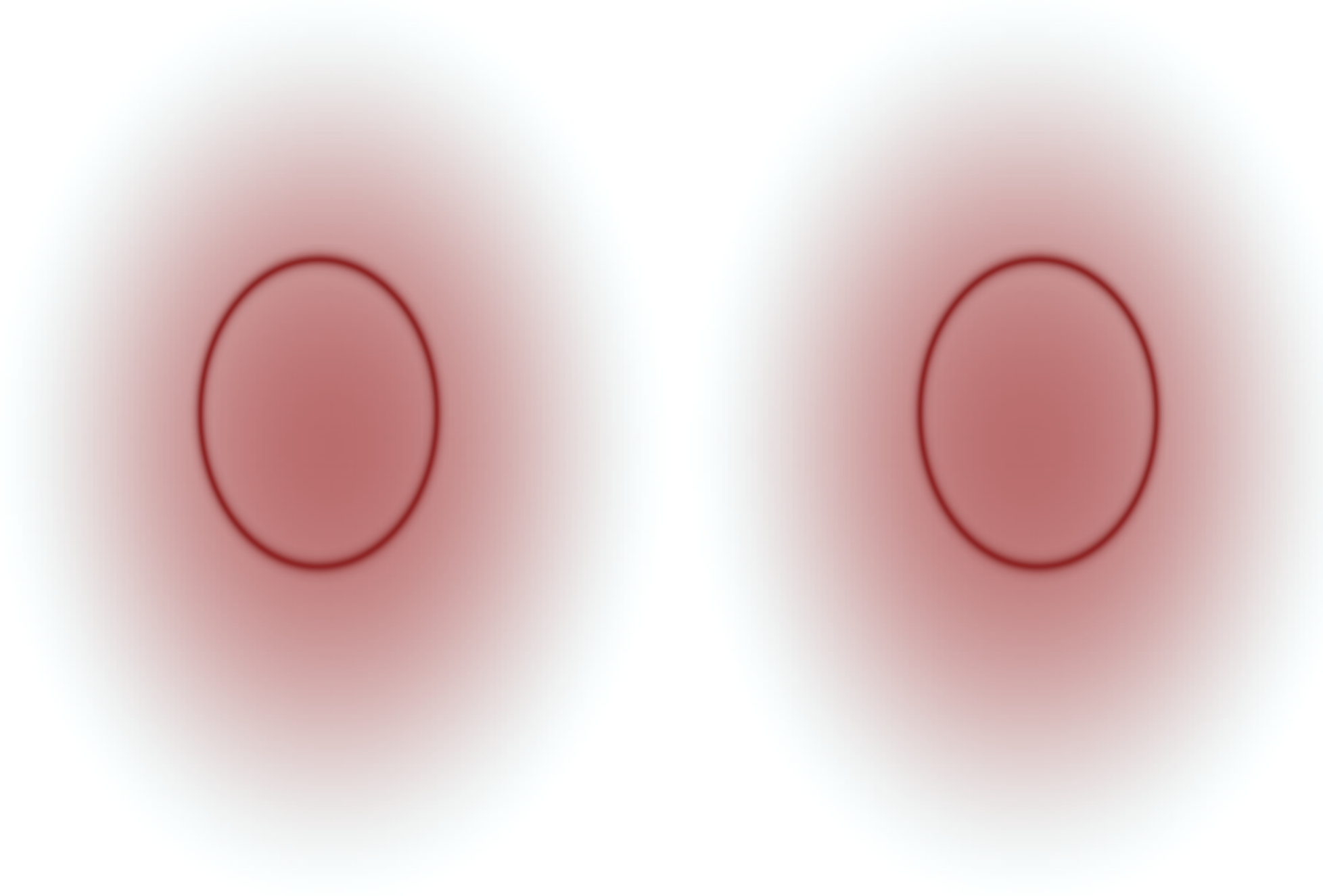
Ultimately, subsampling is effective only in the tall data regime where the data are necessarily redundant.



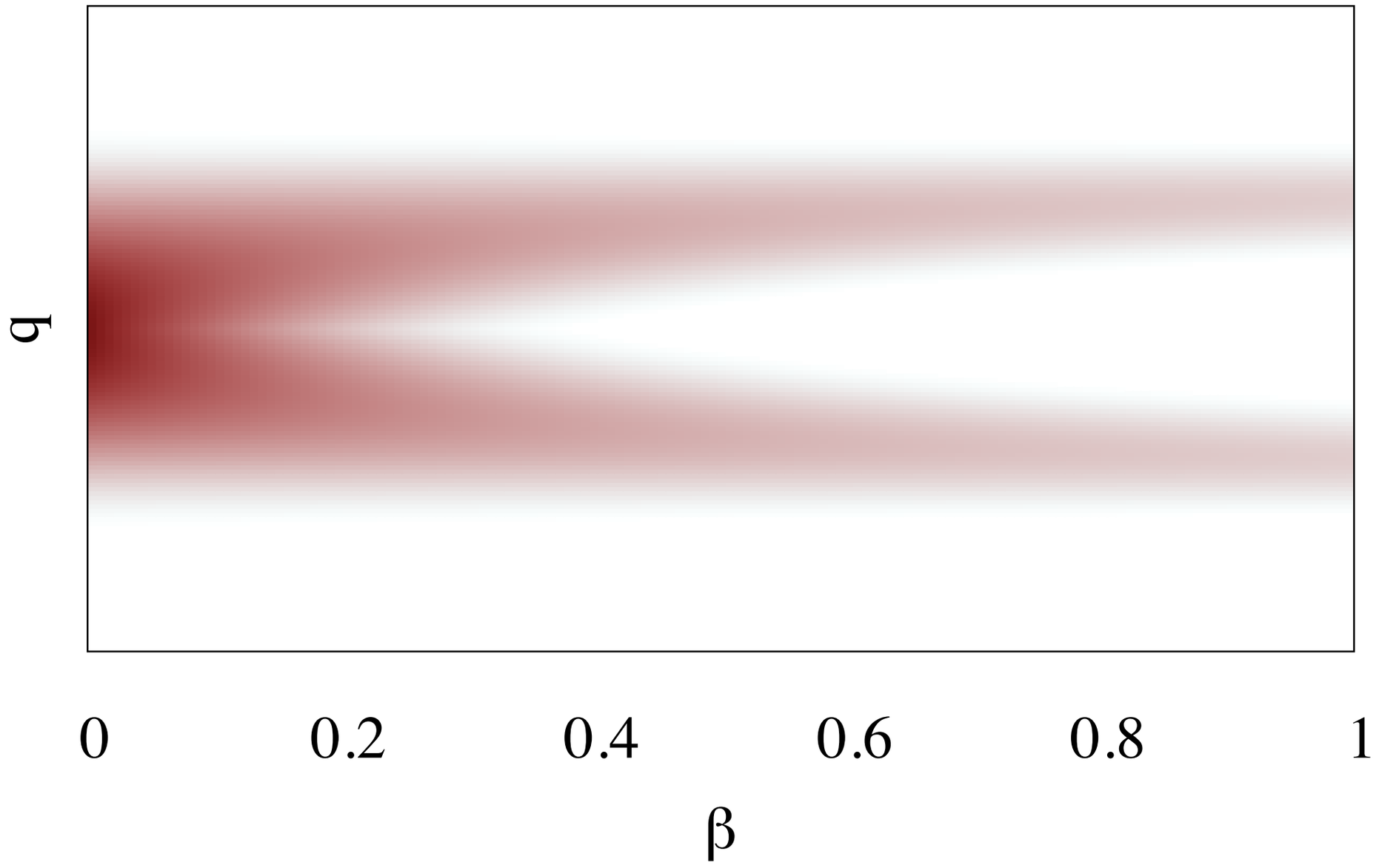
Adiabatic Monte Carlo



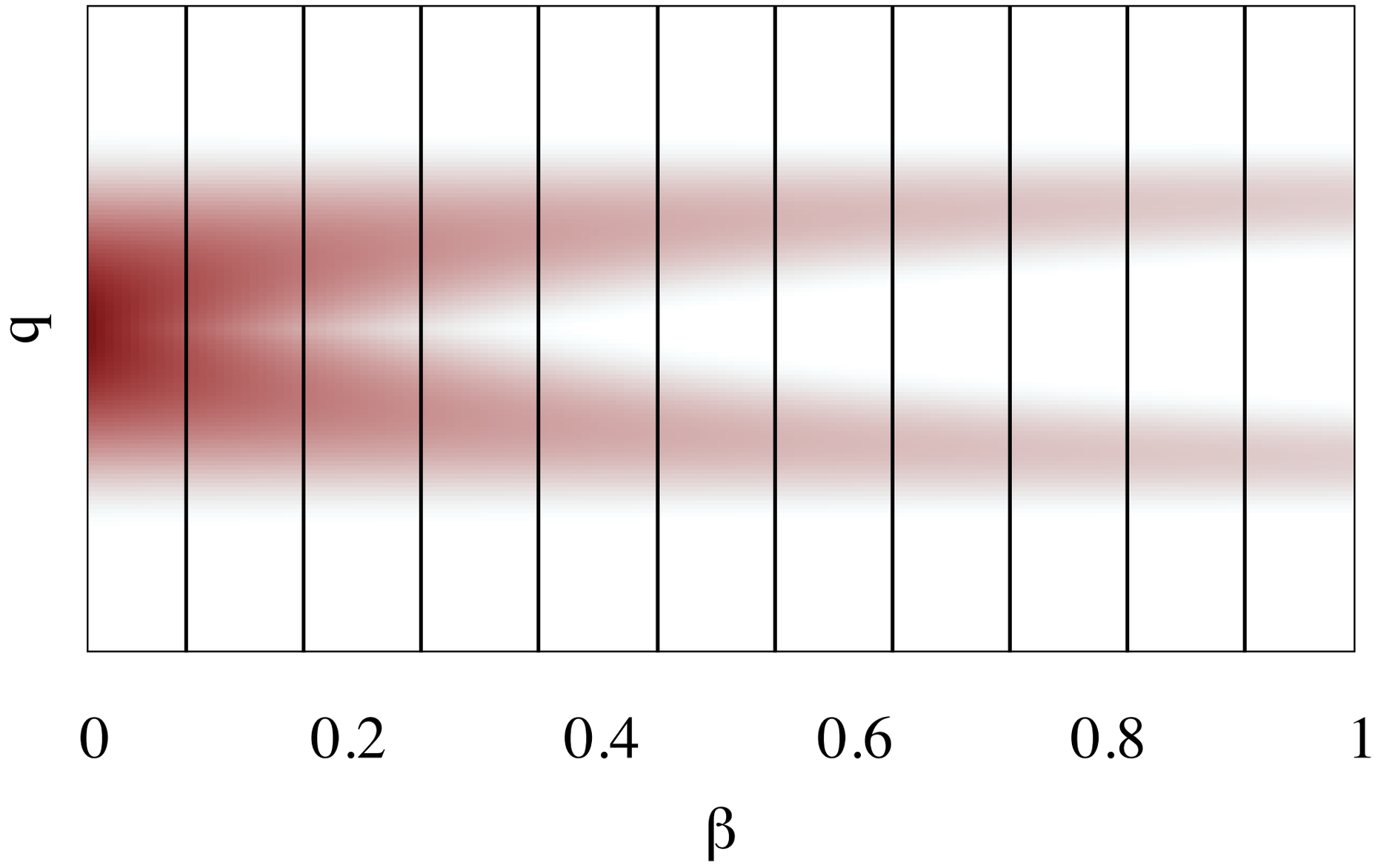
Like any MCMC algorithm, one weakness of Hamiltonian Monte Carlo is multimodal target distributions.



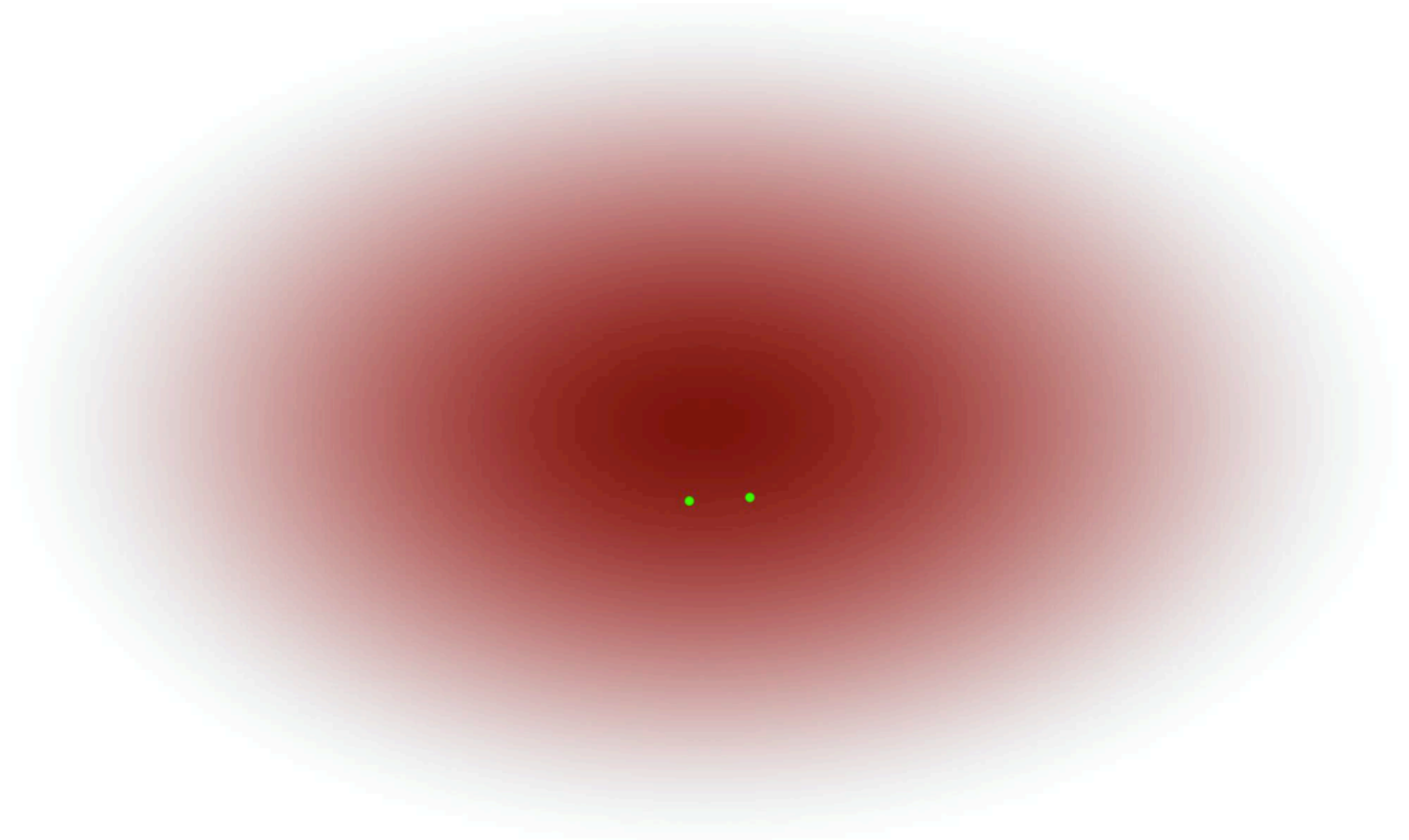
Movement between modes is facilitated by transitioning back and forth to an auxiliary, unimodal distribution.



The efficacy of this approach, however, depends on how efficiently we can transition to this auxiliary distribution.



Adiabatic Monte Carlo uses even more geometry to make these transitions dynamic and avoid bespoke tuning.



Adiabatic Monte Carlo uses even more geometry to make these transitions dynamic and avoid bespoke tuning.