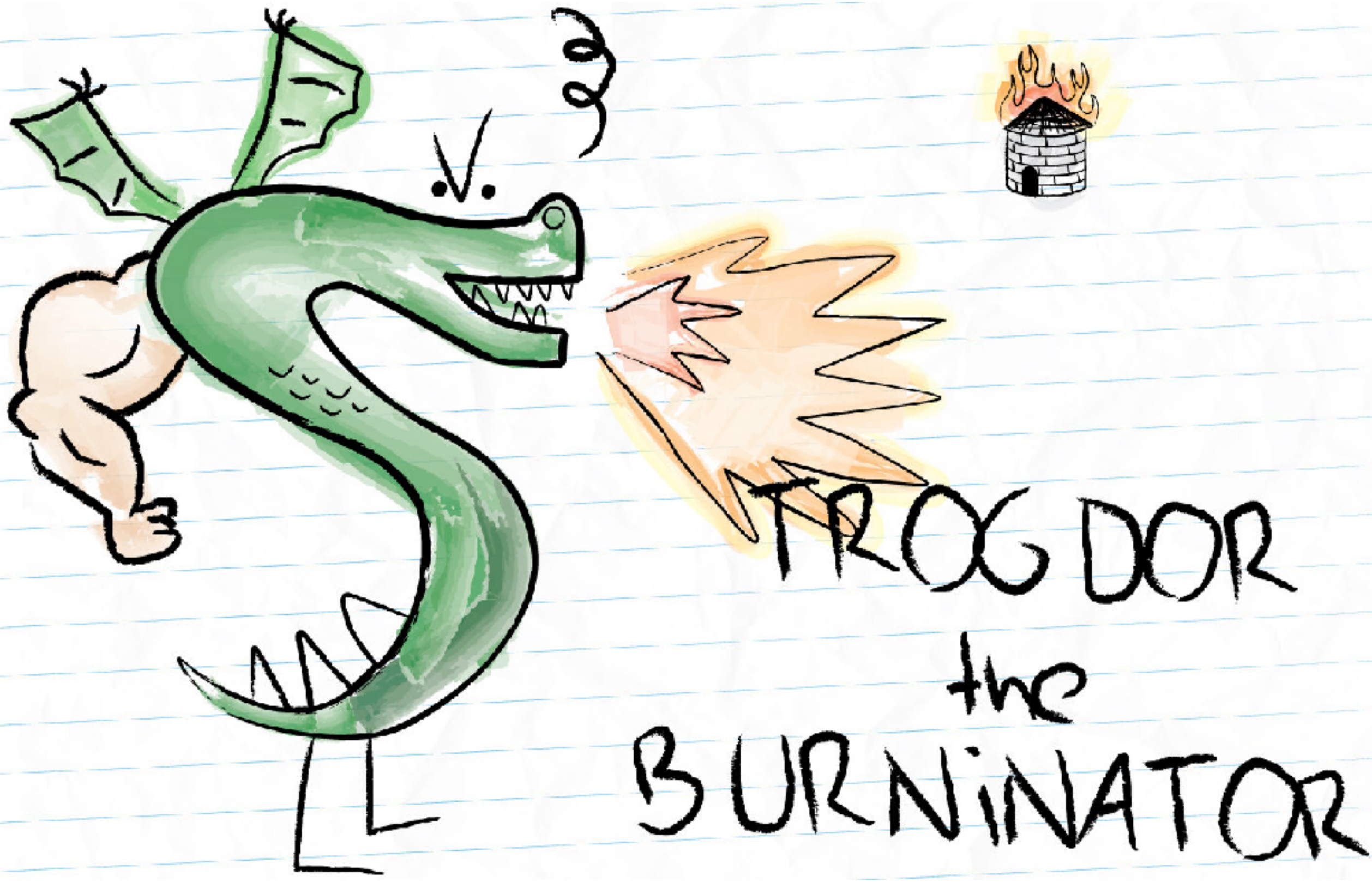
Here Be Hyper-Dragons:
High-Dimensional Spaces and Statistical Computation

Michael Betancourt @betanalpha
Symplectomorphic, LLC

Machine Learning Coffee Seminar,
Aalto University
February 12, 208



https://blitzgraphics.deviantart.com/art/TROGDOR-THE-BURNINATOR-134216353

Regardless of your statistical predilection, all well-posed statistical computations eventually reduce to expectations.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi(\theta) \, f(\theta)$$

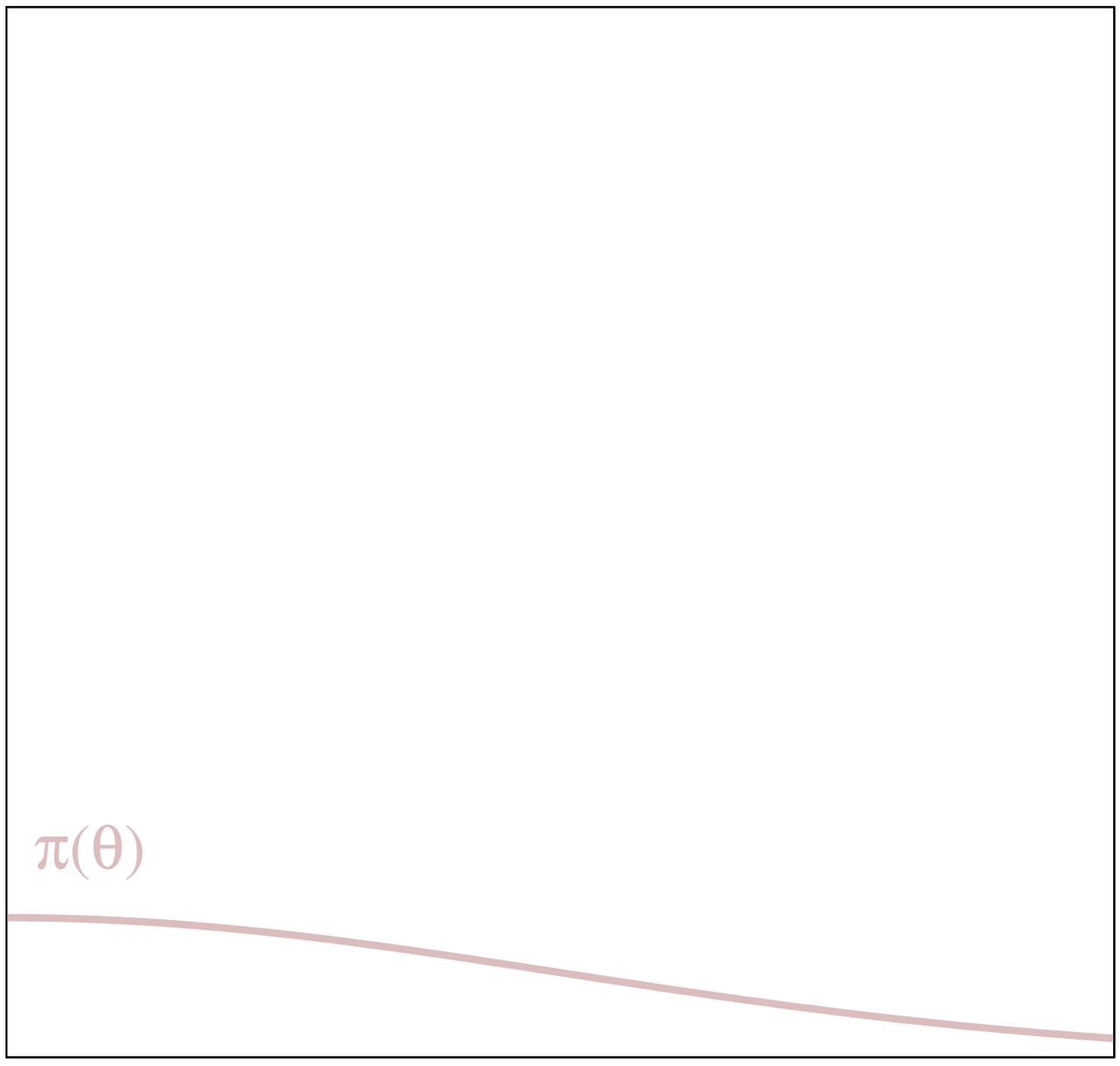
Regardless of your statistical predilection, all well-posed statistical computations eventually reduce to expectations.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi(\theta) \, f(\theta)$$

Regardless of your statistical predilection, all well-posed statistical computations eventually reduce to expectations.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi(\theta) \, f(\theta)$$

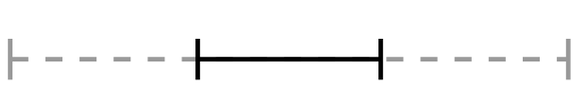
If relevant neighborhoods are determined by *probability density* then we should focus computation near the mode.



But integration doesn't just evaluate the integrand -- it aggregates it over volumes.

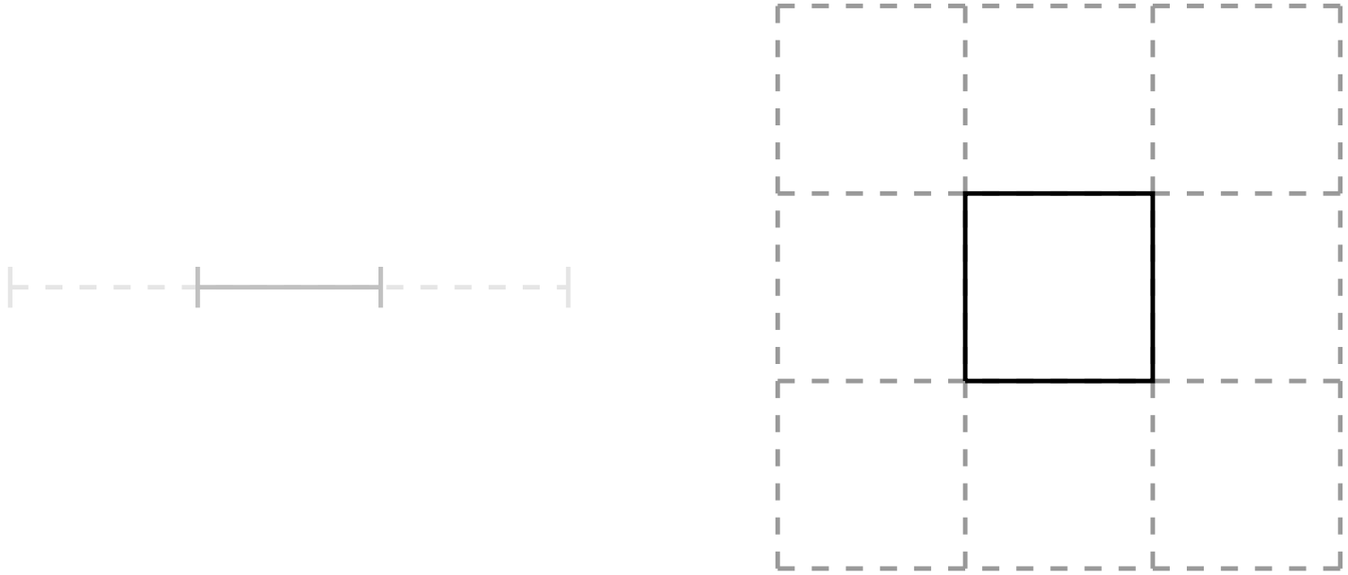
$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi(\theta) \, f(\theta)$$

Volume, however, starts to behave strangely as the dimension of our parameter space increases.



1D

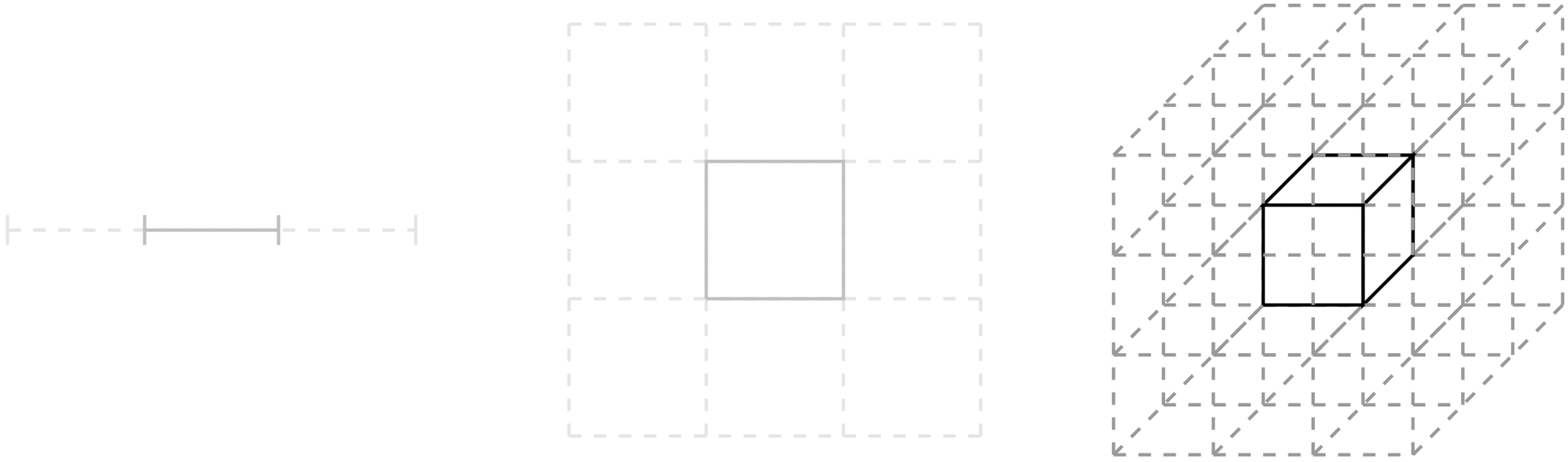
Volume, however, starts to behave strangely as the dimension of our parameter space increases.



1D

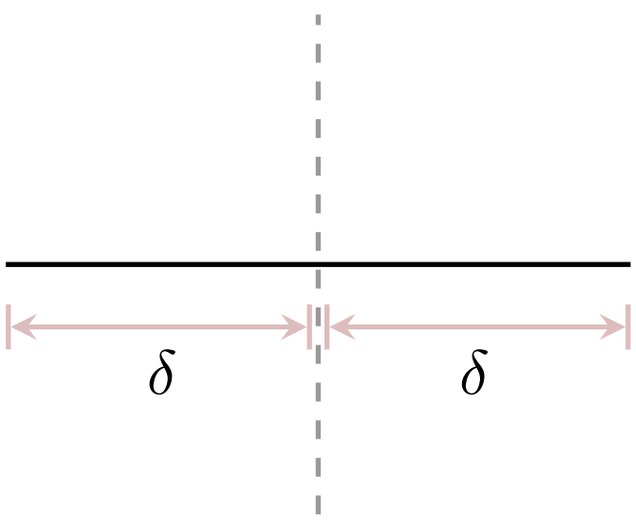
2D

Volume, however, starts to behave strangely as the dimension of our parameter space increases.



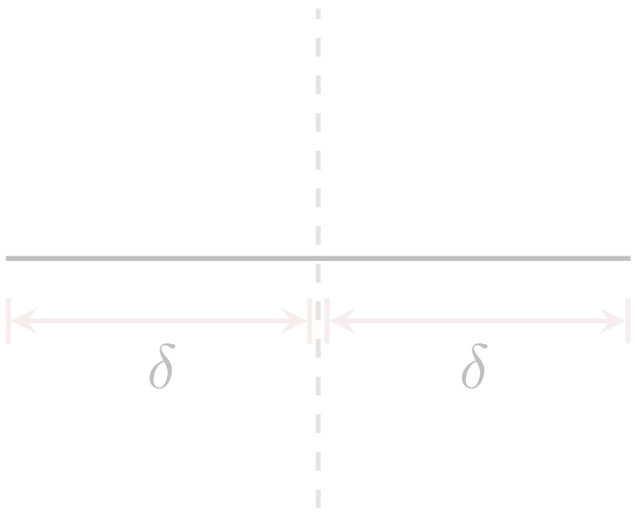
1D 2D 3D

Volume, however, starts to behave strangely as the dimension of our parameter space increases.

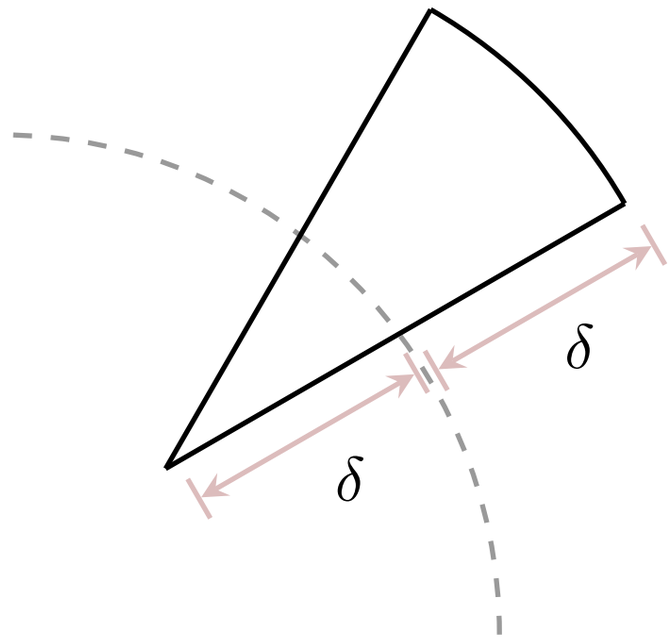


1D

Volume, however, starts to behave strangely as the dimension of our parameter space increases.

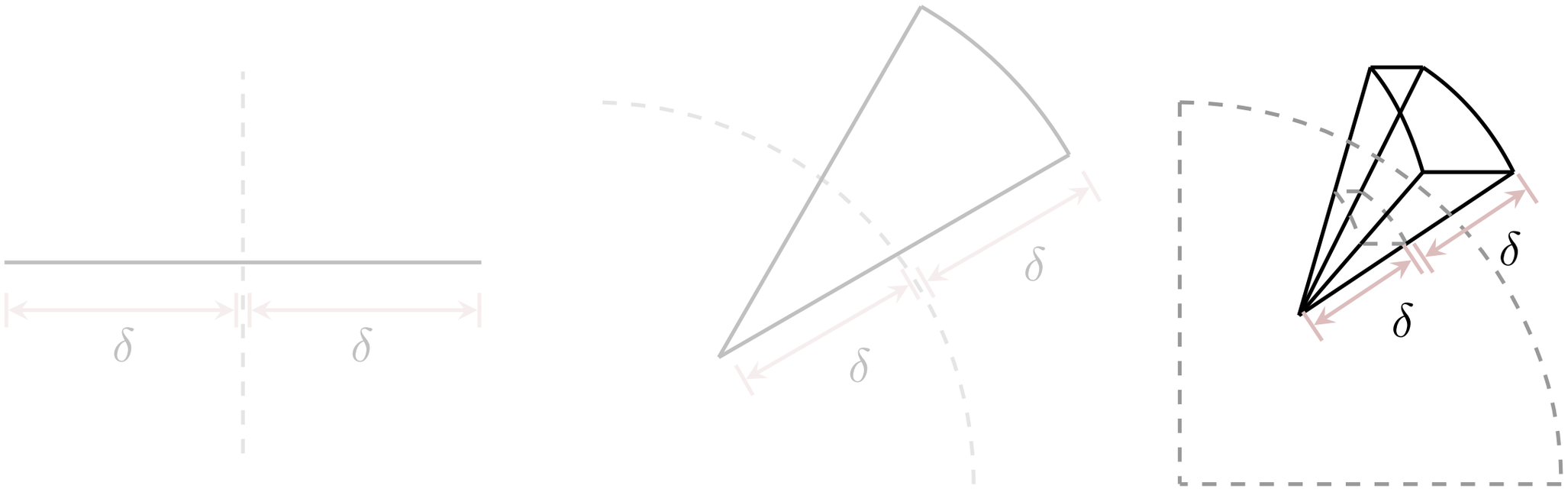


1D



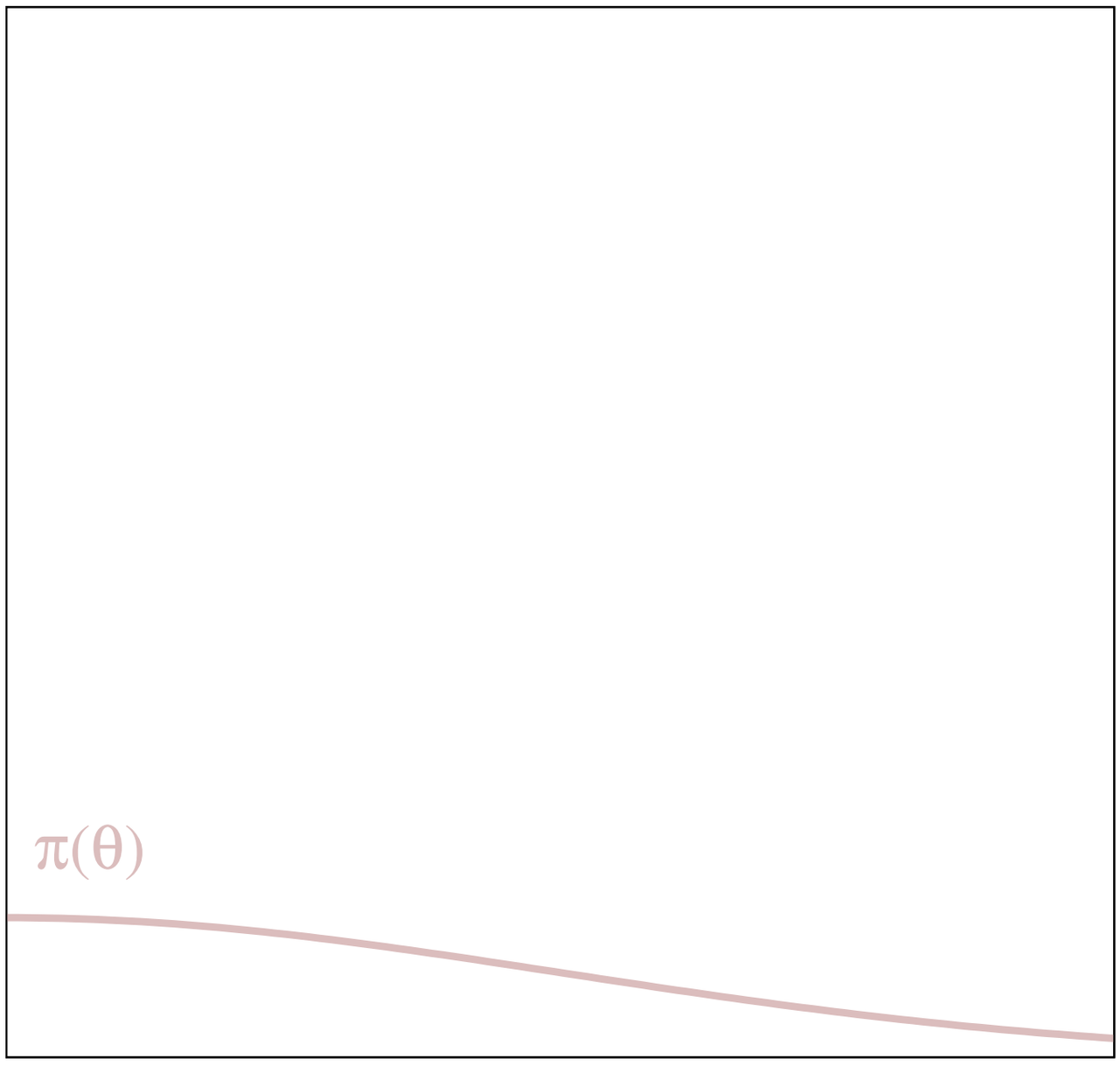
2D

Volume, however, starts to behave strangely as the dimension of our parameter space increases.

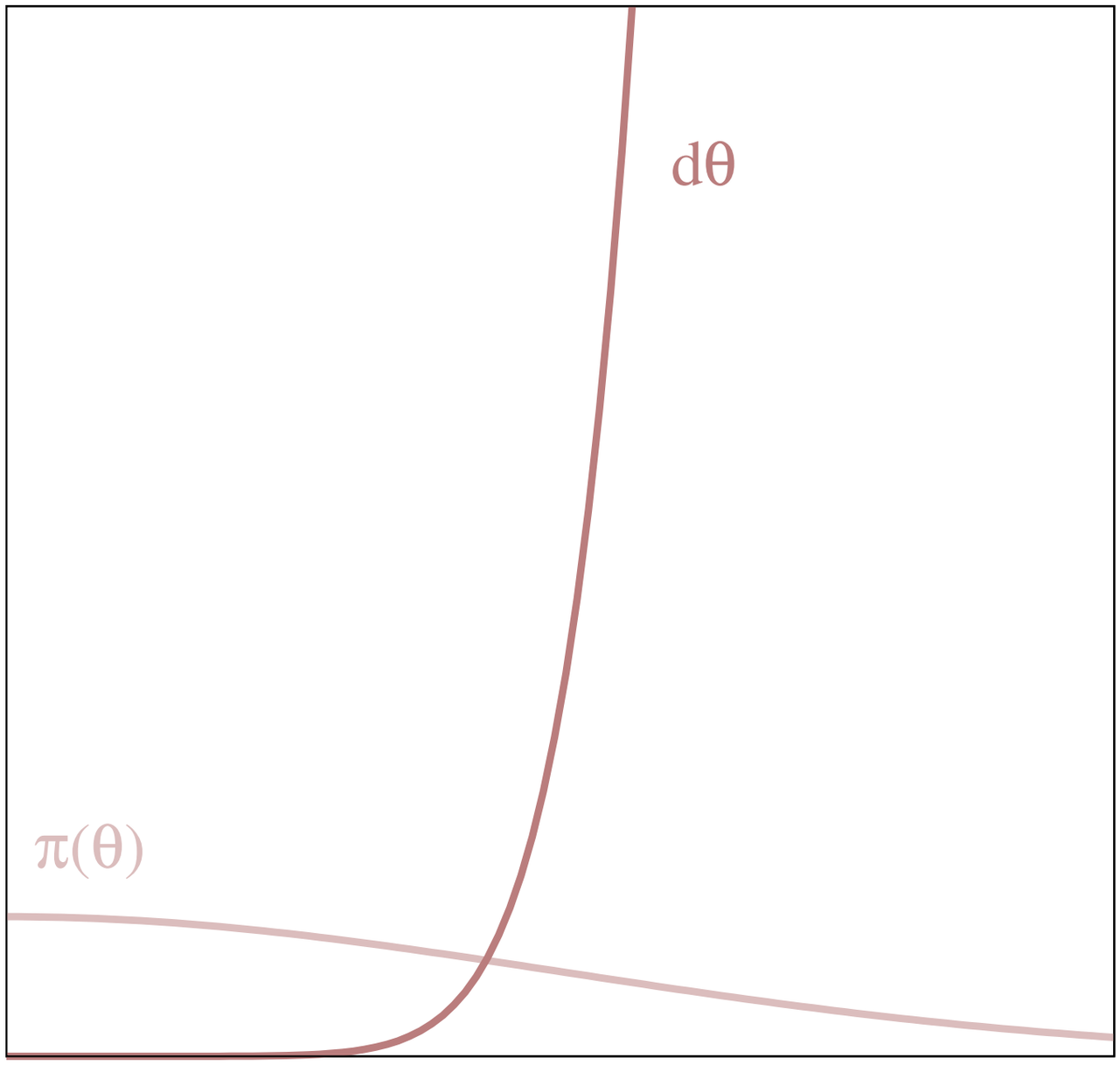


1D 2D 3D

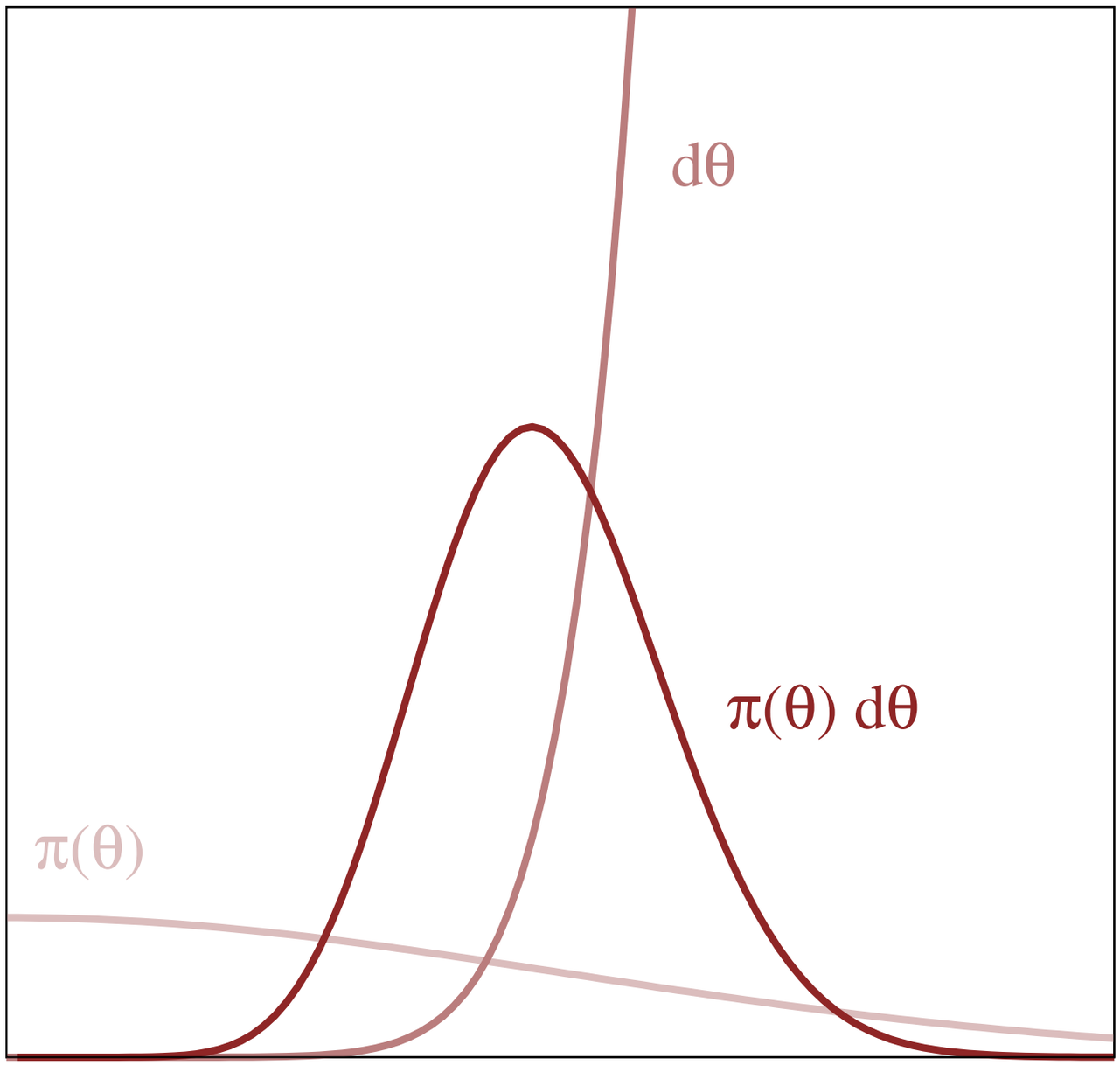
The dominant contributions to an integral is dictated not by probability *density* but rather by probability *mass*.



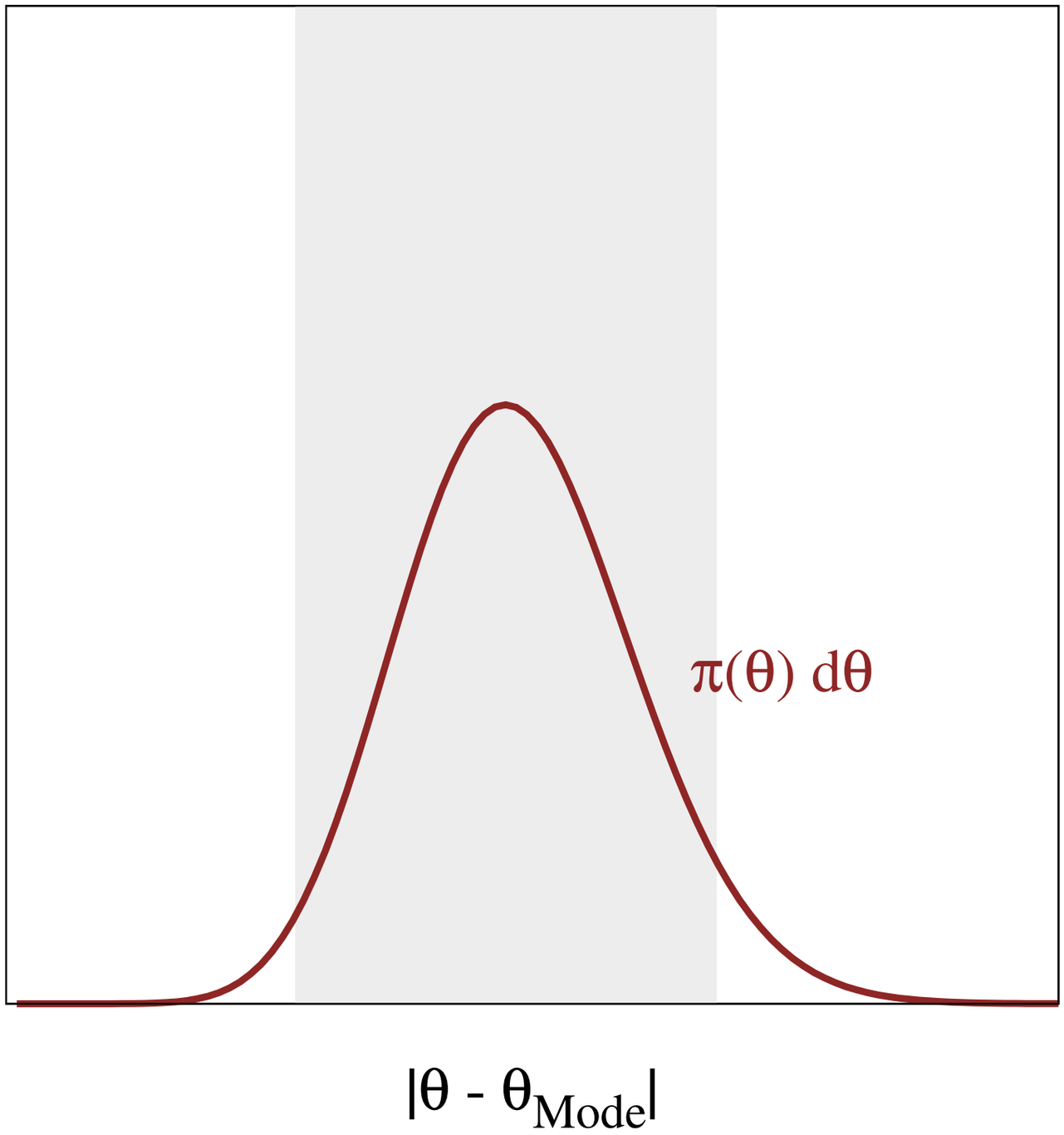
The dominant contributions to an integral is dictated not by probability *density* but rather by probability *mass*.



The dominant contributions to an integral is dictated not by probability *density* but rather by probability *mass*.



As the dimensionality increases, probability mass concentrates near a hypersurface called the *typical set*.



This *concentration of measure* into a nearly singular typical set frustrates the accurate estimation of integrals.

In order to accurately approximating expectations computational methods must quantify the typical set.

Deterministic

Modal Estimators

Laplace Estimators

Variational Estimators

...

Stochastic

Rejection Sampling

Importance Sampling

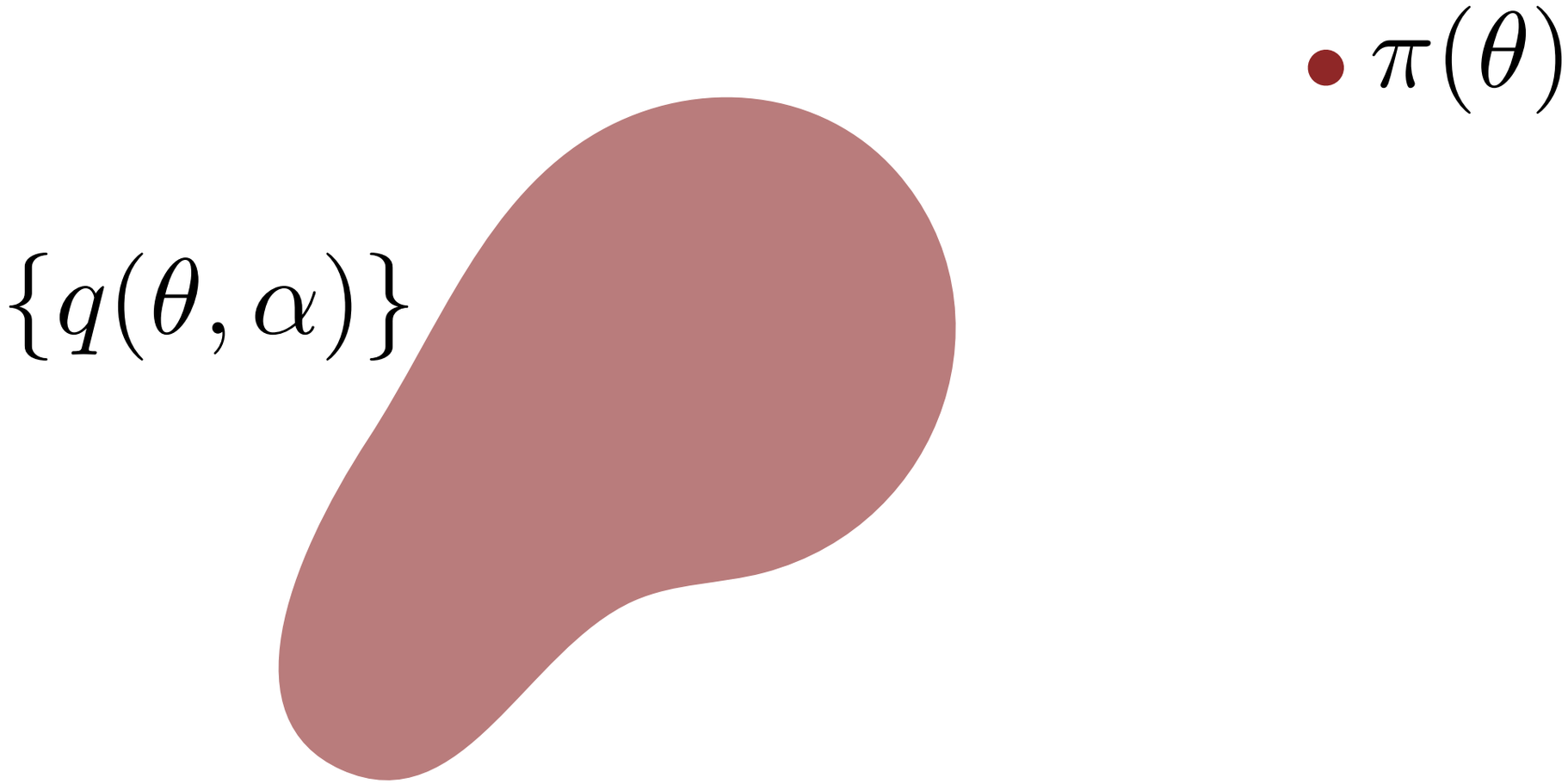
Markov Chain Monte Carlo

...

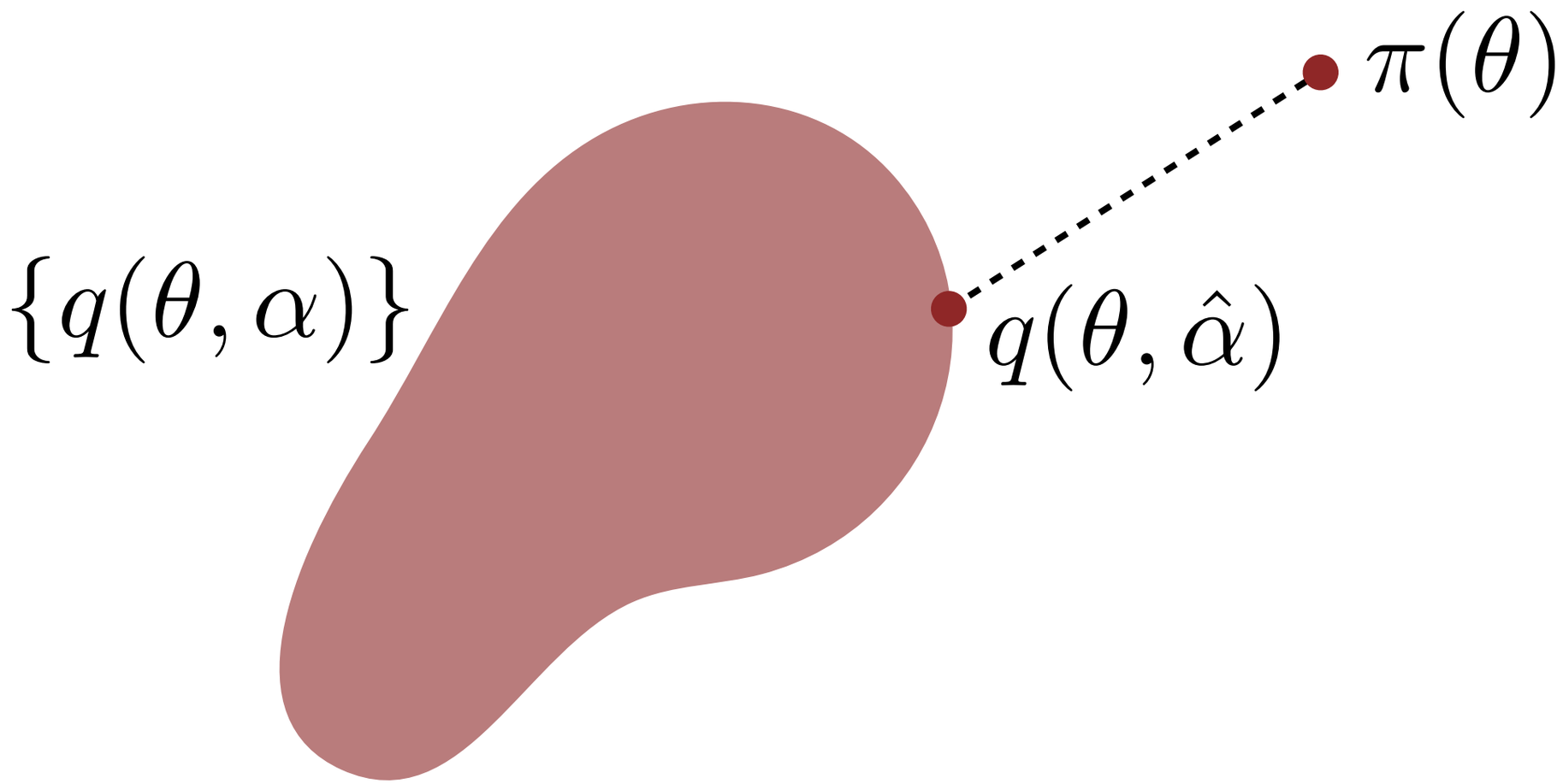
Variational methods optimize over a family of convenient approximating distributions.

- $\pi(\theta)$

Variational methods optimize over a family of convenient approximating distributions.



Variational methods optimize over a family of convenient approximating distributions.



The (local) variational solution is then used to approximate the target expectations.

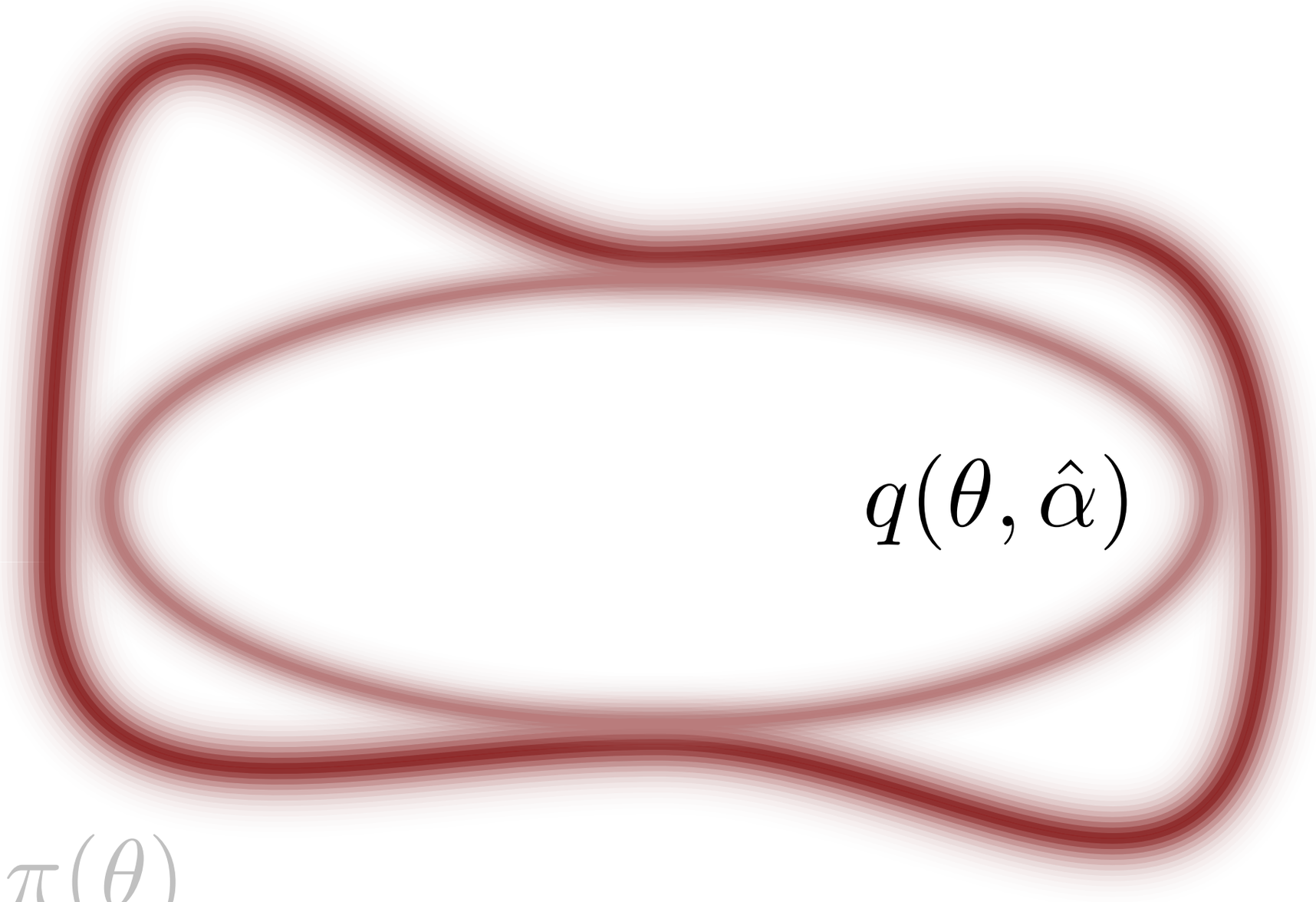
$$\pi(\theta) \approx q(\theta, \hat{\alpha})$$

$$\int f(\theta)\, \pi(\theta)\, \mathrm{d}\theta \approx \int f(\theta)\, q(\theta, \hat{\alpha})\, \mathrm{d}\theta$$

The (local) variational solution is then used to approximate the target expectations.

$\pi(\theta)$

The (local) variational solution is then used to approximate the target expectations.



The (local) variational solution is then used to approximate the target expectations.

$\pi(\theta)$

The (local) variational solution is then used to approximate the target expectations.

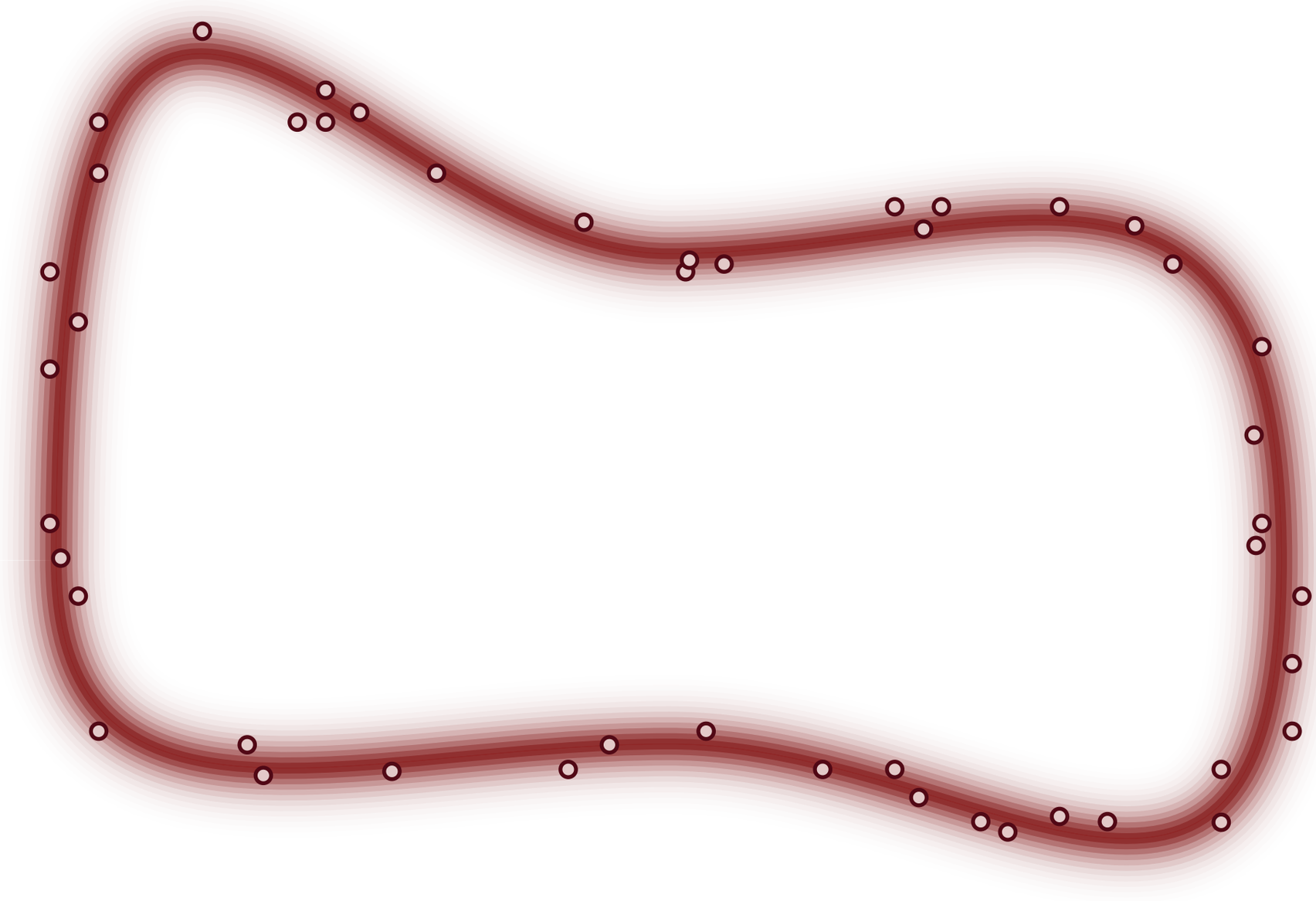
Stochastic methods construct estimators that converge to the exact target expectations.

$$\{\theta_1, \ldots, \theta_N\}$$

$$\int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta \approx \frac{\sum_{n=1}^{N} w(\theta_n)\,f(\theta_n)}{\sum_{n=1}^{N} w(\theta_n)}$$

Monte Carlo methods quantify the typical set using exact samples drawn from the target distribution.

Monte Carlo methods quantify the typical set using exact samples drawn from the target distribution.



Monte Carlo estimators average a given function over these samples to approximate the expectation.

$$\frac{1}{N}\sum_{n=1}^{N} f(\theta_n) \sim \mathcal{N}\left(\mathbb{E}[f], \frac{\mathrm{Var}[f]}{N}\right)$$

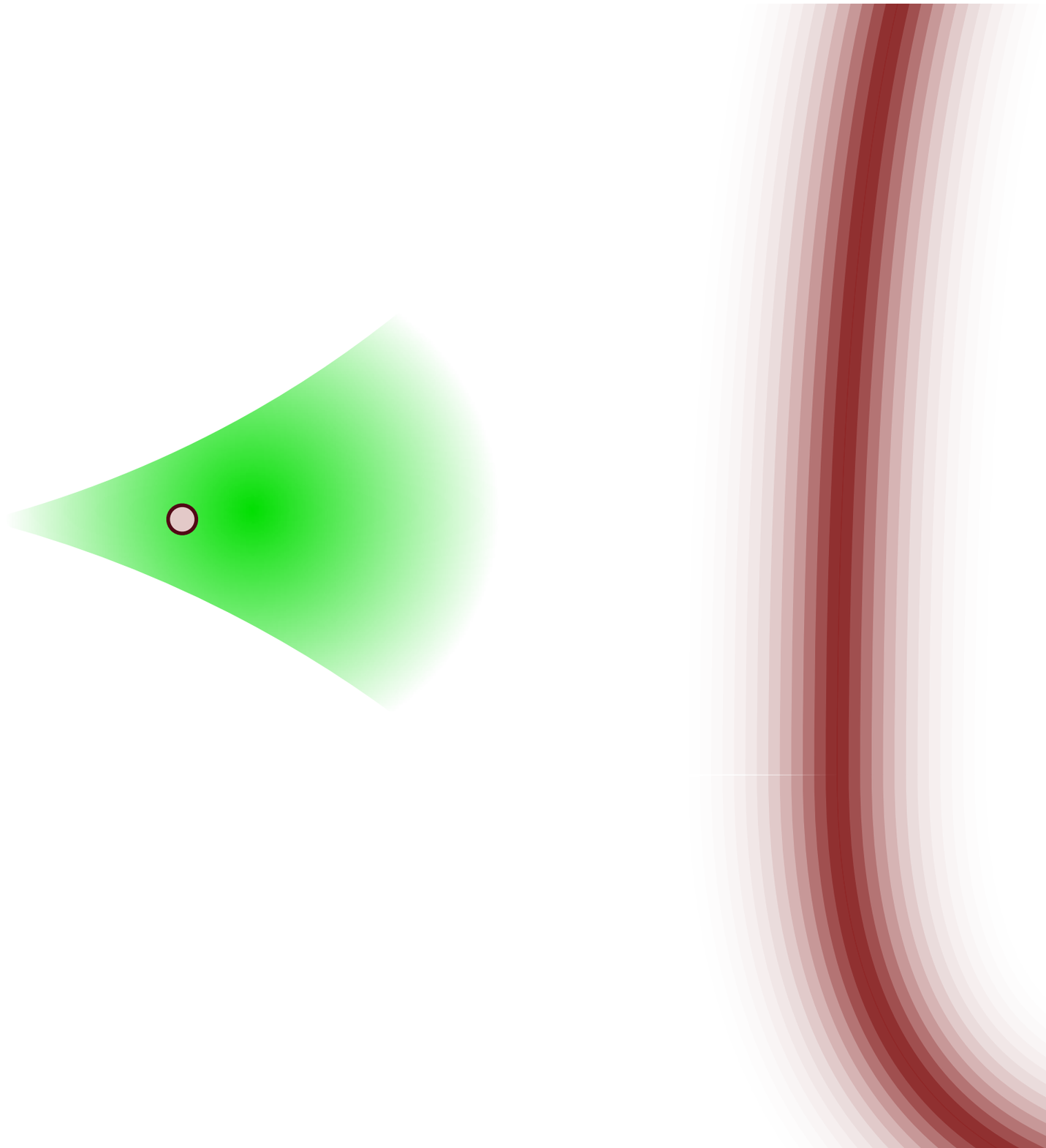
Usually we can't generate exact samples from complex target distributions, but we can generate *correlated* samples.

$$T(\theta \mid \theta')$$

Usually we can't generate exact samples from complex target distributions, but we can generate *correlated* samples.

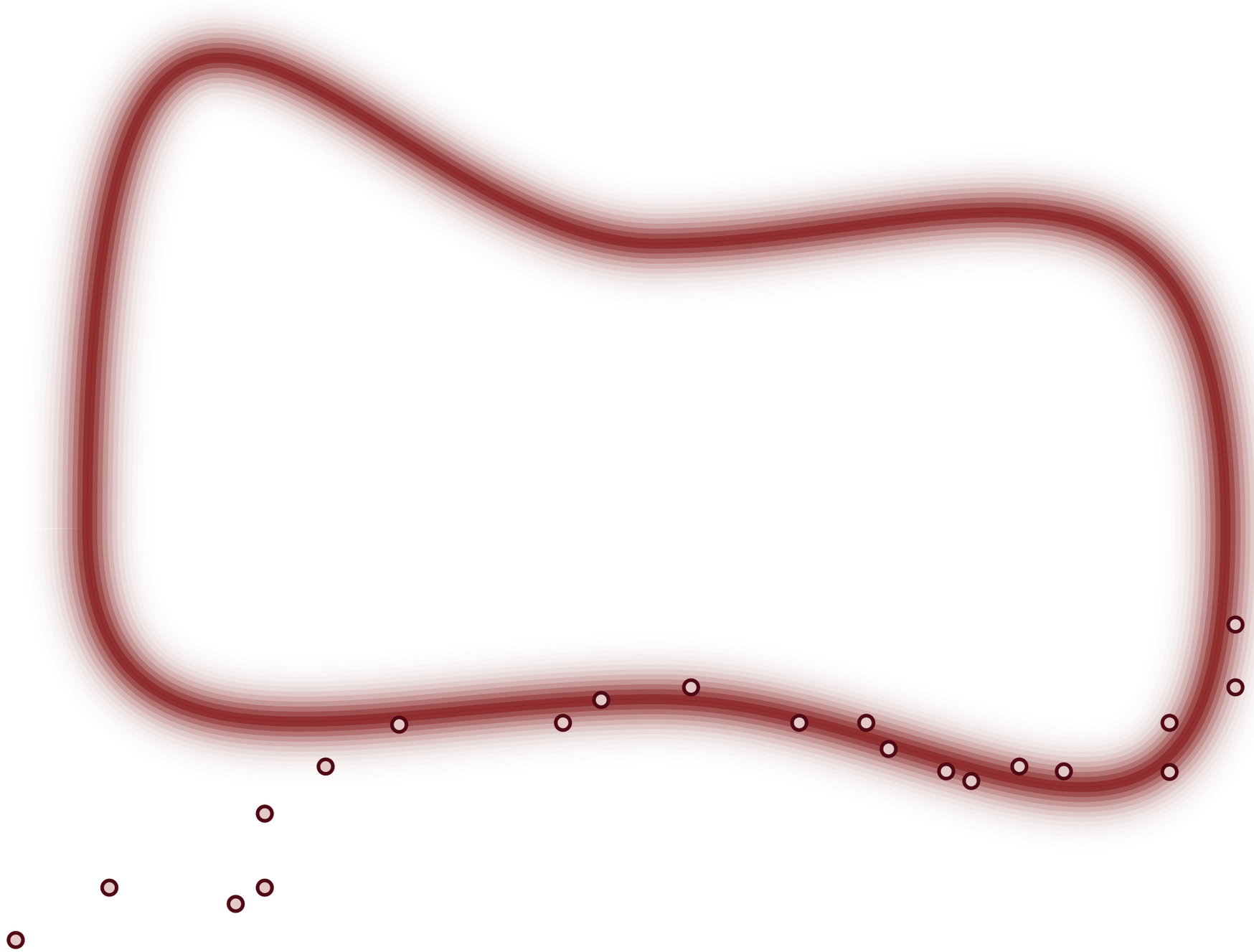
$$\pi(\theta) = \int T(\theta \mid \theta') \, \pi(\theta') \, \mathrm{d}\theta'$$

A Markov transition that preserves the target distribution naturally concentrates towards the typical set.

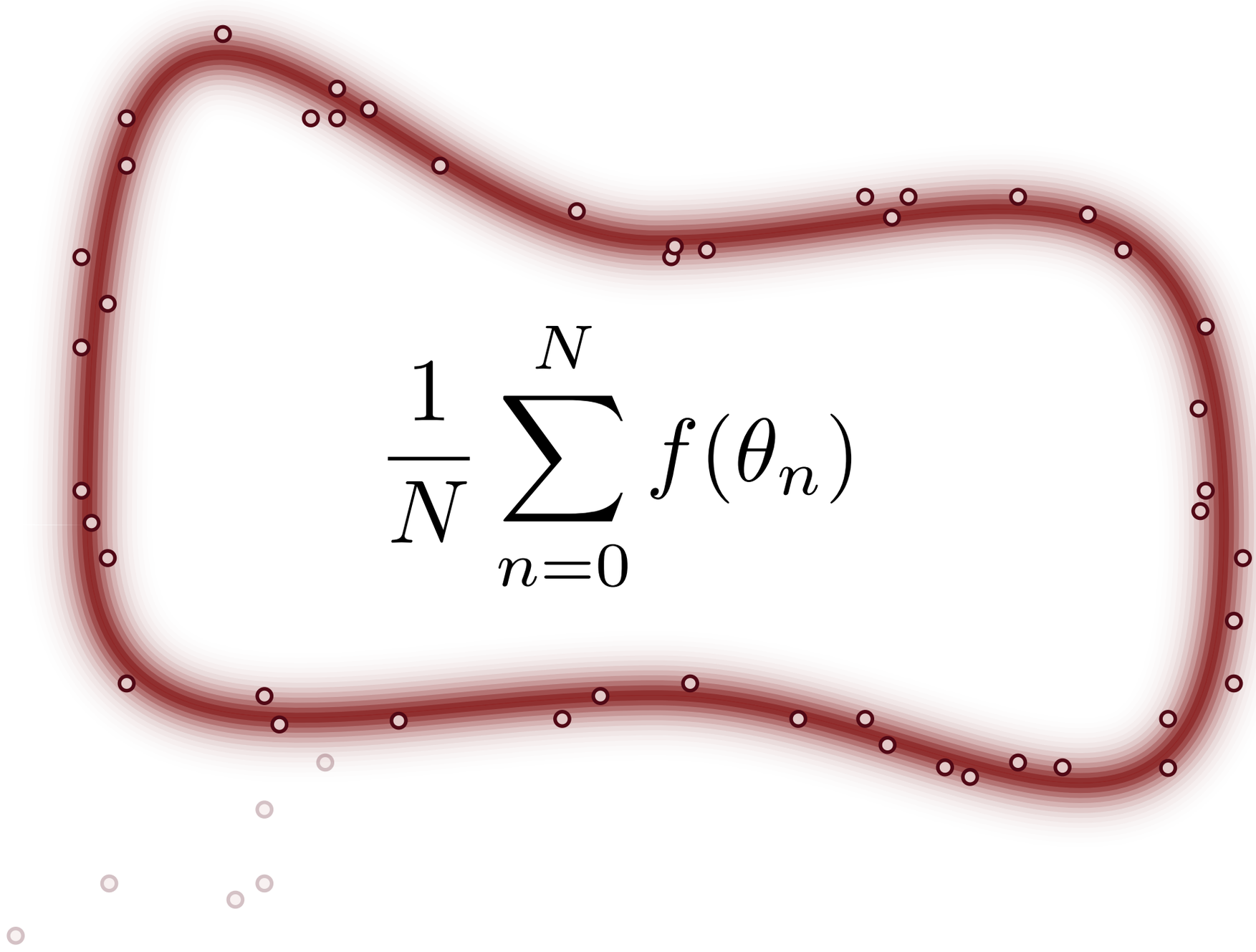


Markov chains then provide a generic scheme for finding and then exploring the typical set.

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

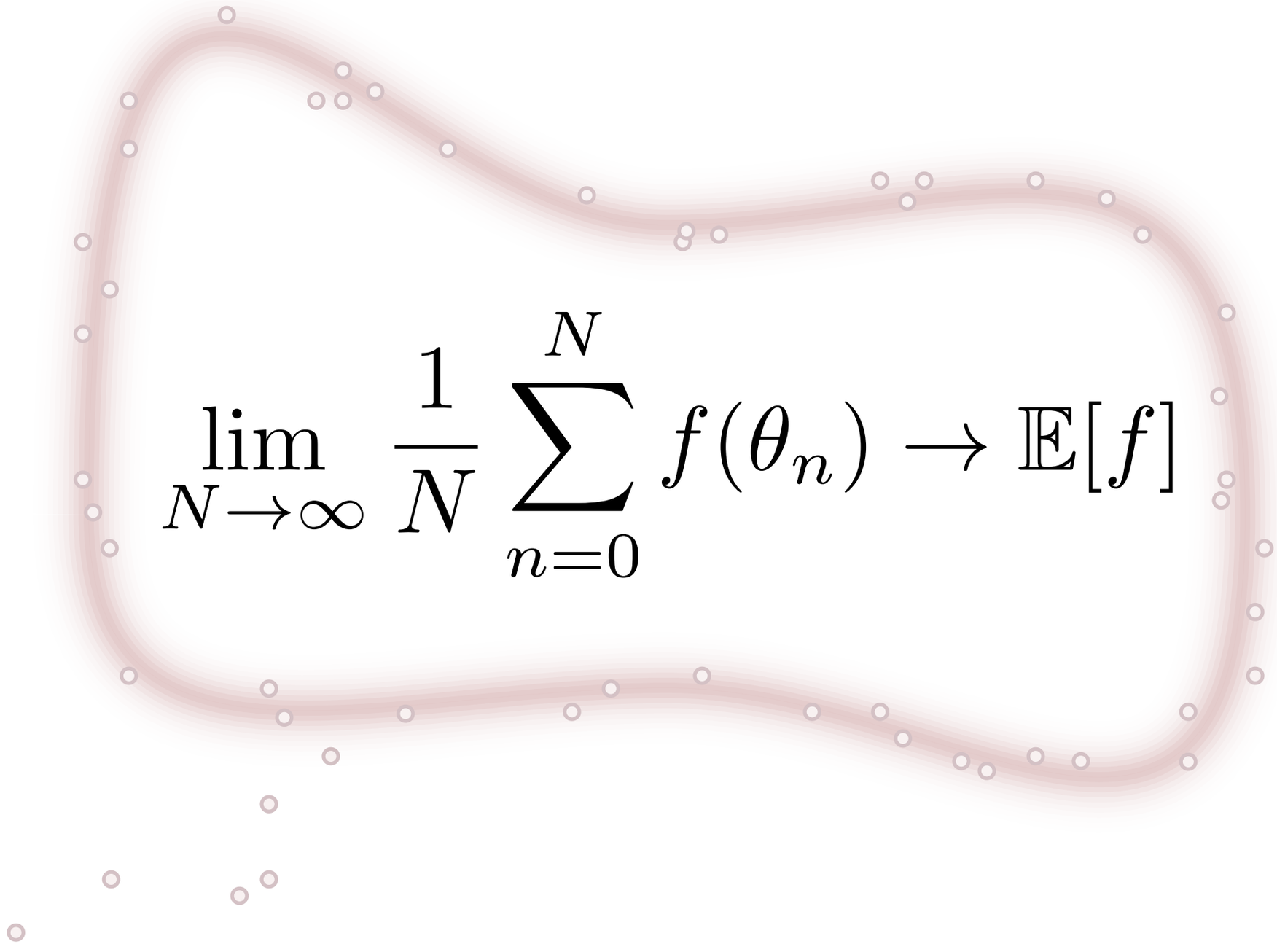


If run long the chain long enough then we can construct consistent *Markov Chain Monte Carlo estimators*.



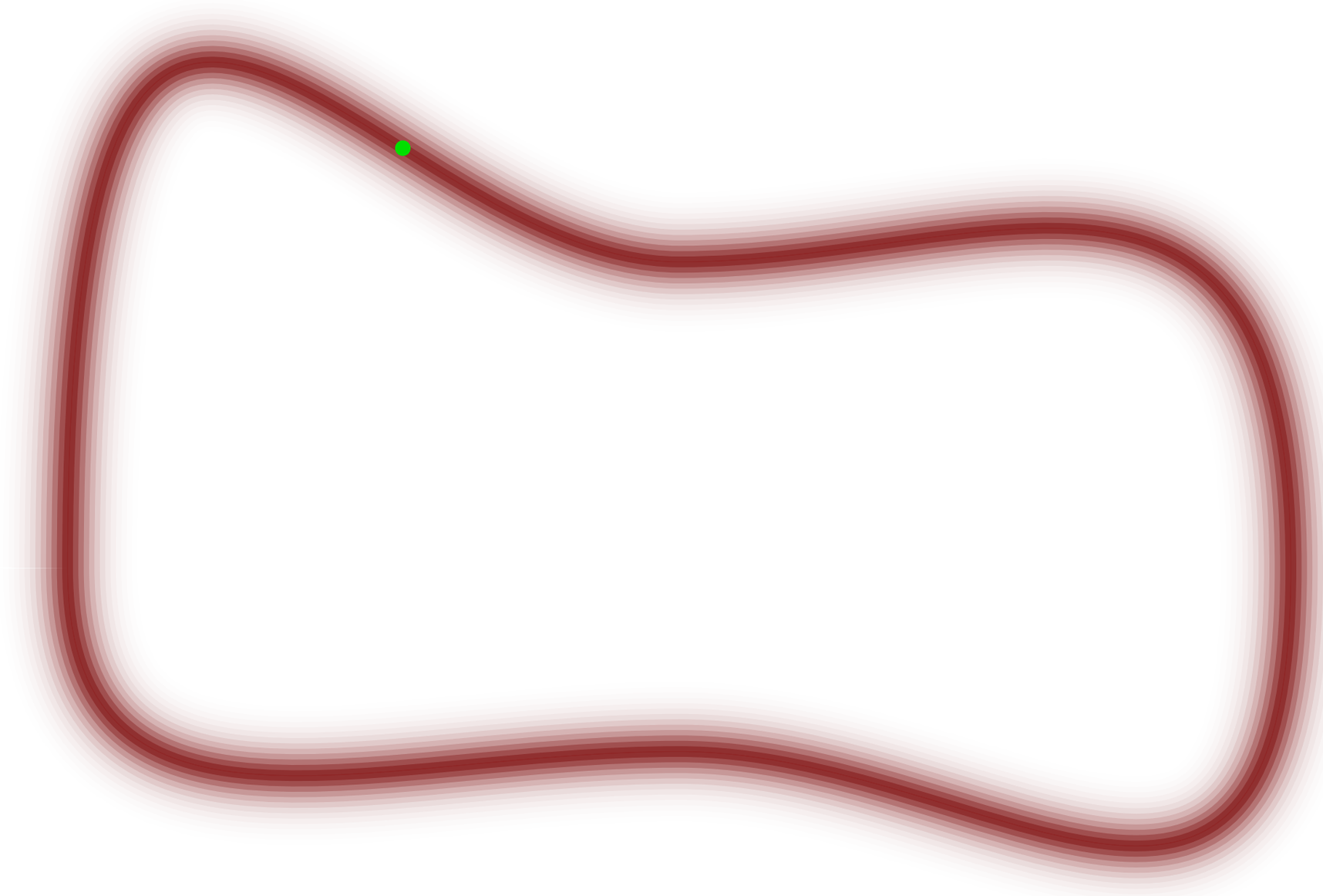
$$\frac{1}{N}\sum_{n=0}^{N} f(\theta_n)$$

If run long the chain long enough then we can construct consistent *Markov Chain Monte Carlo estimators*.

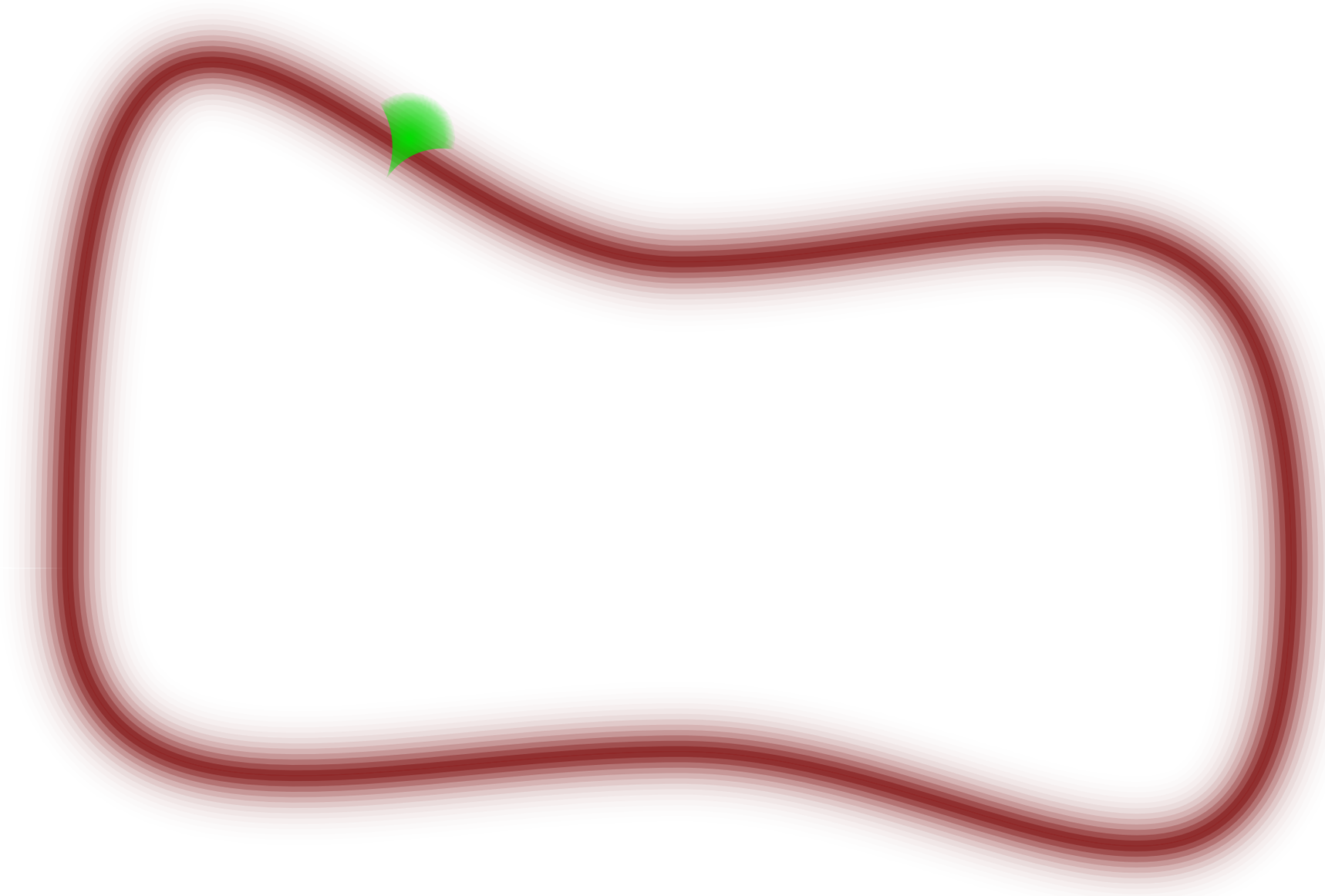


$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=0}^{N} f(\theta_n) \to \mathbb{E}[f]$$

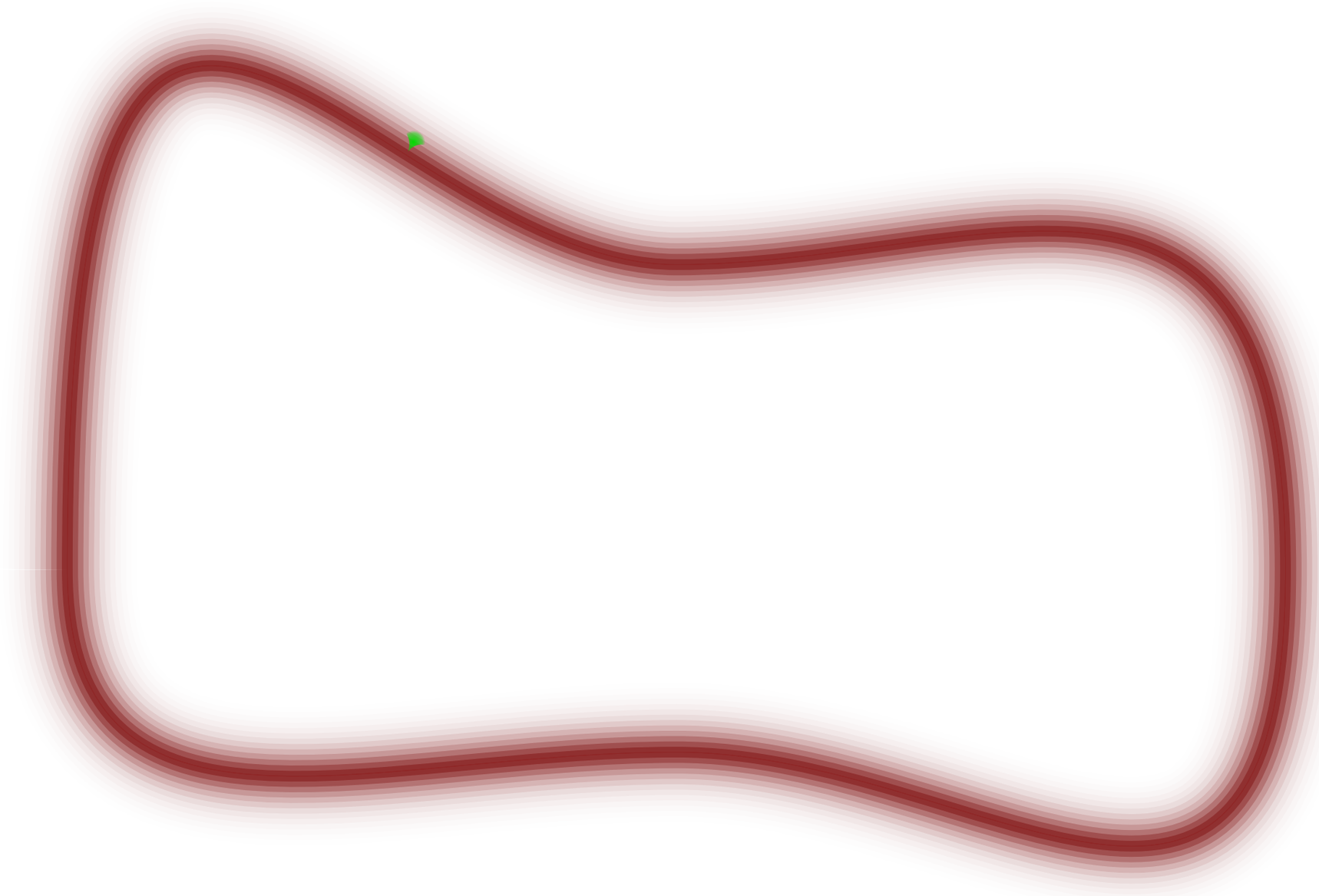
Unfortunately the performance of simple algorithms like Random Walk Metropolis does not scale well.



Unfortunately the performance of simple algorithms like Random Walk Metropolis does not scale well.



Unfortunately the performance of simple algorithms like Random Walk Metropolis does not scale well.



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.

Approximations, such as data subsampling, perturb the typical set and frustrate accurate computation.

Approximations, such as data subsampling, perturb the typical set and frustrate accurate computation.

Approximations, such as data subsampling, perturb the typical set and frustrate accurate computation.

Always be wary of approximations that require overlapping typical sets in high dimensions.

Always be wary of approximations that require overlapping typical sets in high dimensions.

Always be wary of approximations that require overlapping typical sets in high dimensions.

Always be wary of approximations that require overlapping typical sets in high dimensions.