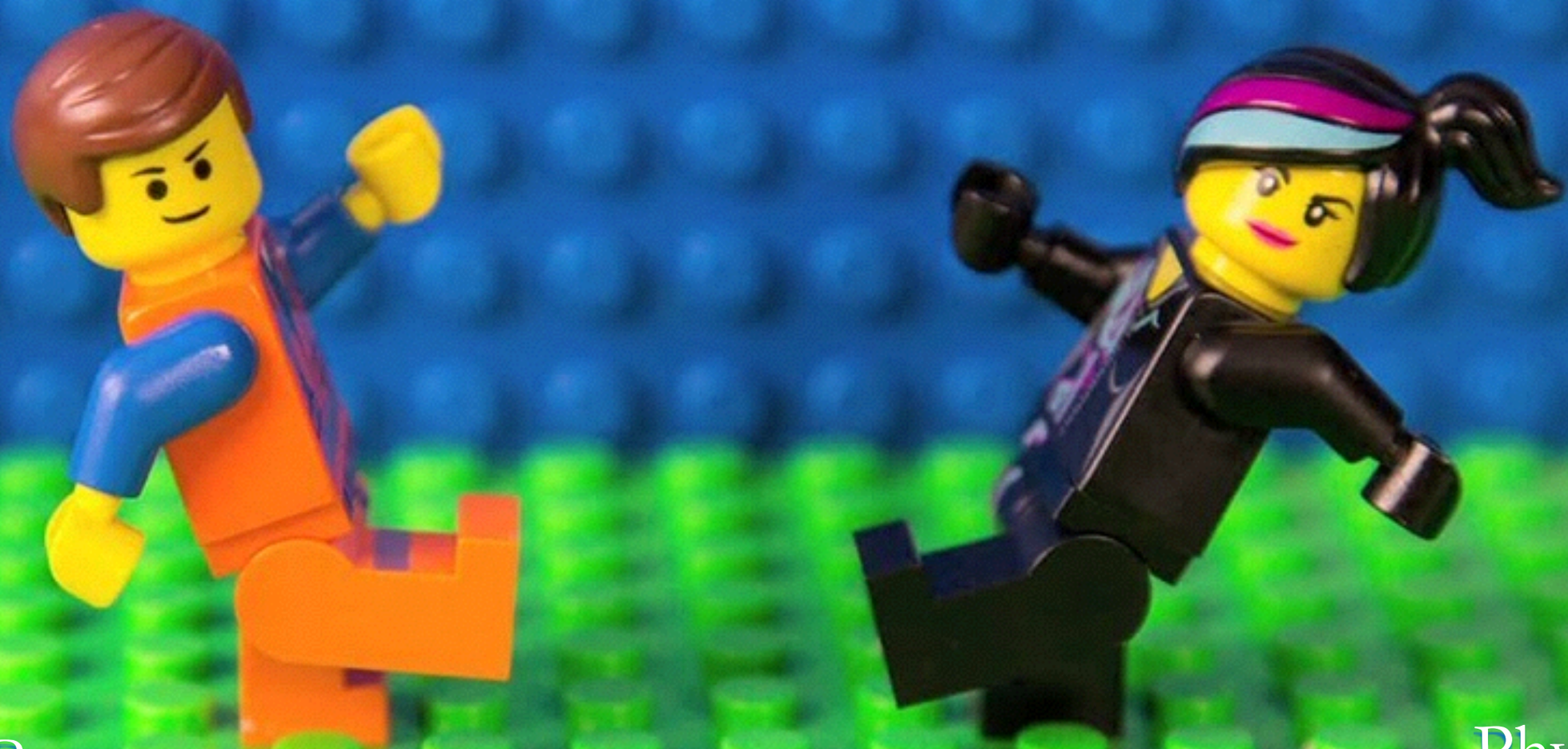
EVERYTHING IS AWESOME

AN INTEGRAL

Fundamentals of Statistical Computing



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

PhyStat-υ

The Kavli Institute for the Physics
and Mathematics of the Universe,
Kashiwa, Japan

May 31, 2016

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

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

Frequentist methods compute expectations with respect to the data to identify estimators that work well *on average*.

$$\hat{\theta}(\mathcal{D})$$

Frequentist methods compute expectations with respect to the data to identify estimators that work well *on average*.

$$\hat{\theta}(\mathcal{D})$$

$$\mathcal{L}(\hat{\theta}(\mathcal{D}), \theta)$$

Frequentist methods compute expectations with respect to the data to identify estimators that work well *on average*.

$$\hat{\theta}(\mathcal{D})$$

$$\mathcal{L}(\theta) = \int \mathcal{L}(\hat{\theta}(\mathcal{D}) \, , \theta) \pi(\mathcal{D} \mid \theta) \, \mathrm{d}\mathcal{D}$$

Bayesian methods compute expectations with respect to the posterior to quantify quantify uncertainty.

$$\pi(\theta \mid \mathcal{D}) \propto \pi(\mathcal{D} \mid \theta)\, \pi(\theta)$$

Bayesian methods compute expectations with respect to the posterior to quantify quantify uncertainty.

$$\pi(\theta \mid \mathcal{D}) \propto \pi(\mathcal{D} \mid \theta)\, \pi(\theta)$$

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

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

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

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

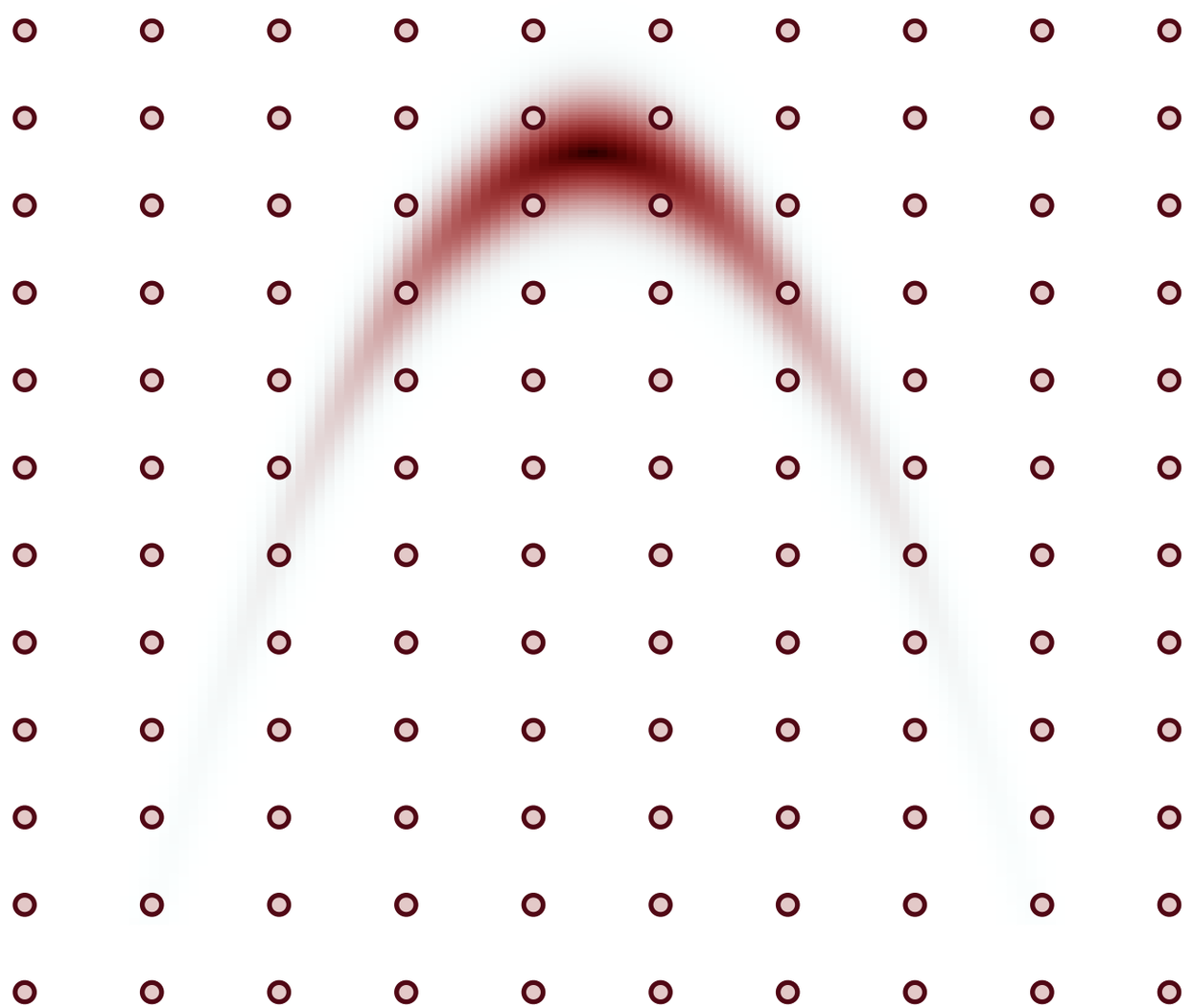
$$\mathbb{E}[f] = \int f(\theta)\, \pi(\theta)\, \mathrm{d}\theta$$

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

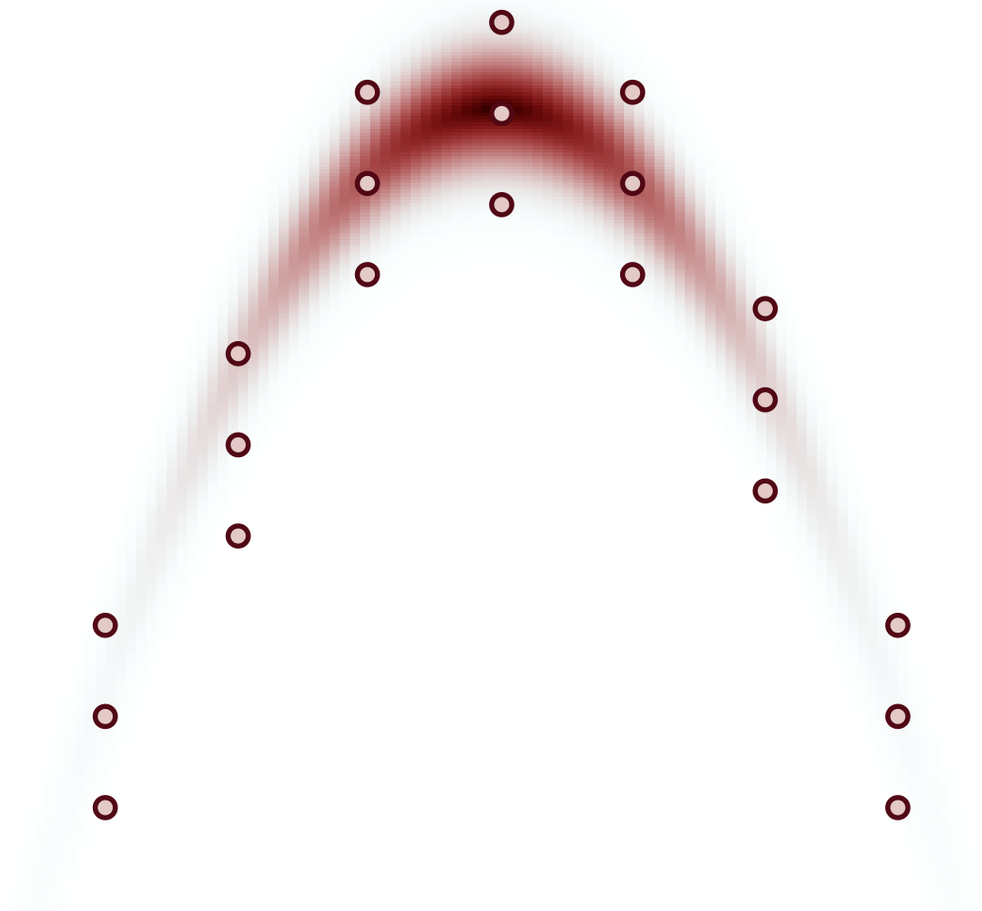
$$\mathbb{E}[f] = \int f(\theta)\, \pi(\theta)\, \mathrm{d}\theta$$

The cost of naive algorithms, like numerical quadrature, scales exponentially with dimension.

The cost of naive algorithms, like numerical quadrature, scales exponentially with dimension.



To scale any better we need to avoid wasteful computation with a more optimally-designed grid.



The ideal grid, however, is somewhat counterintuitive
-- it has to adapt to probability *mass*, not density.

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

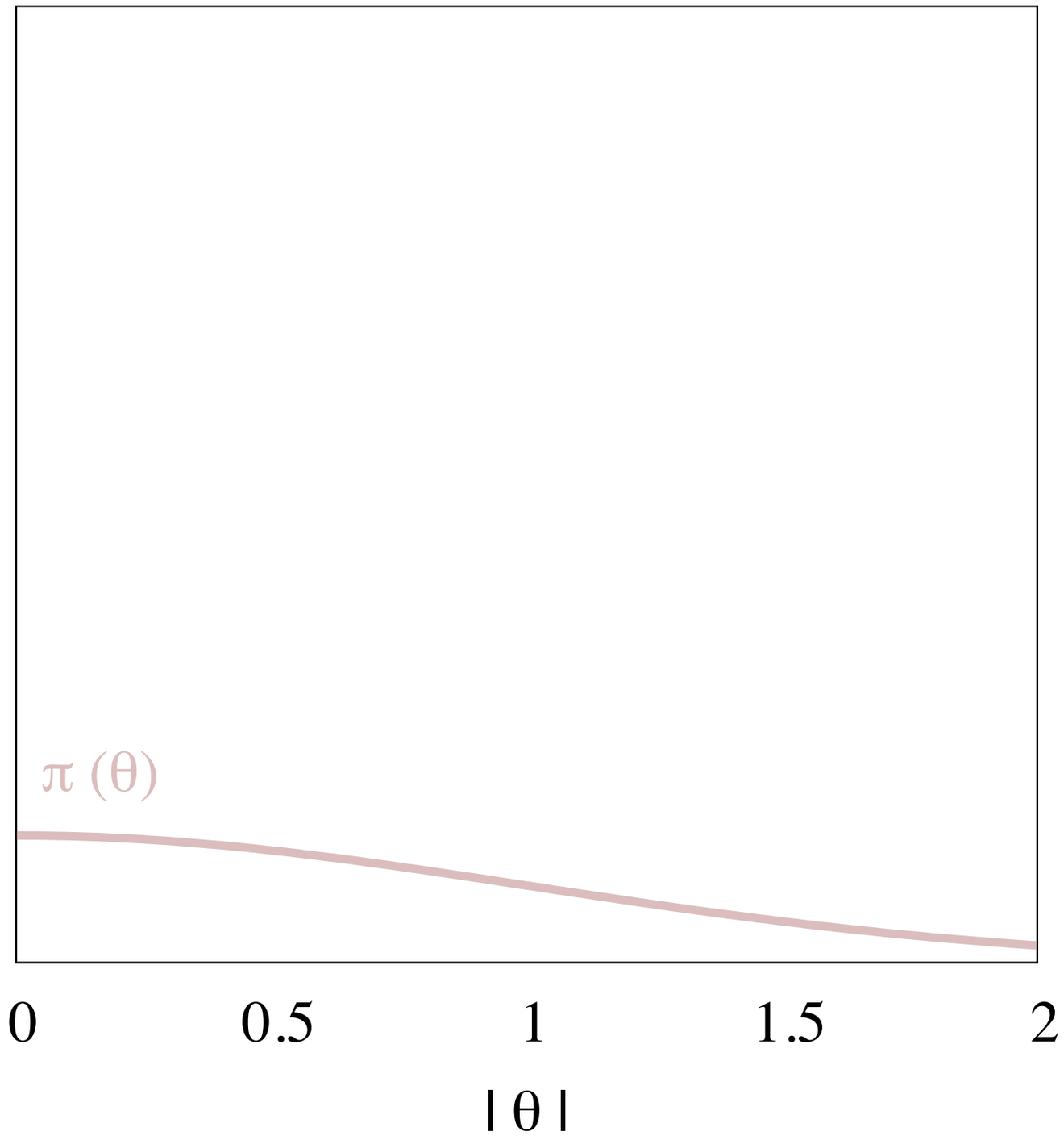
The ideal grid, however, is somewhat counterintuitive
-- it has to adapt to probability *mass*, not density.

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

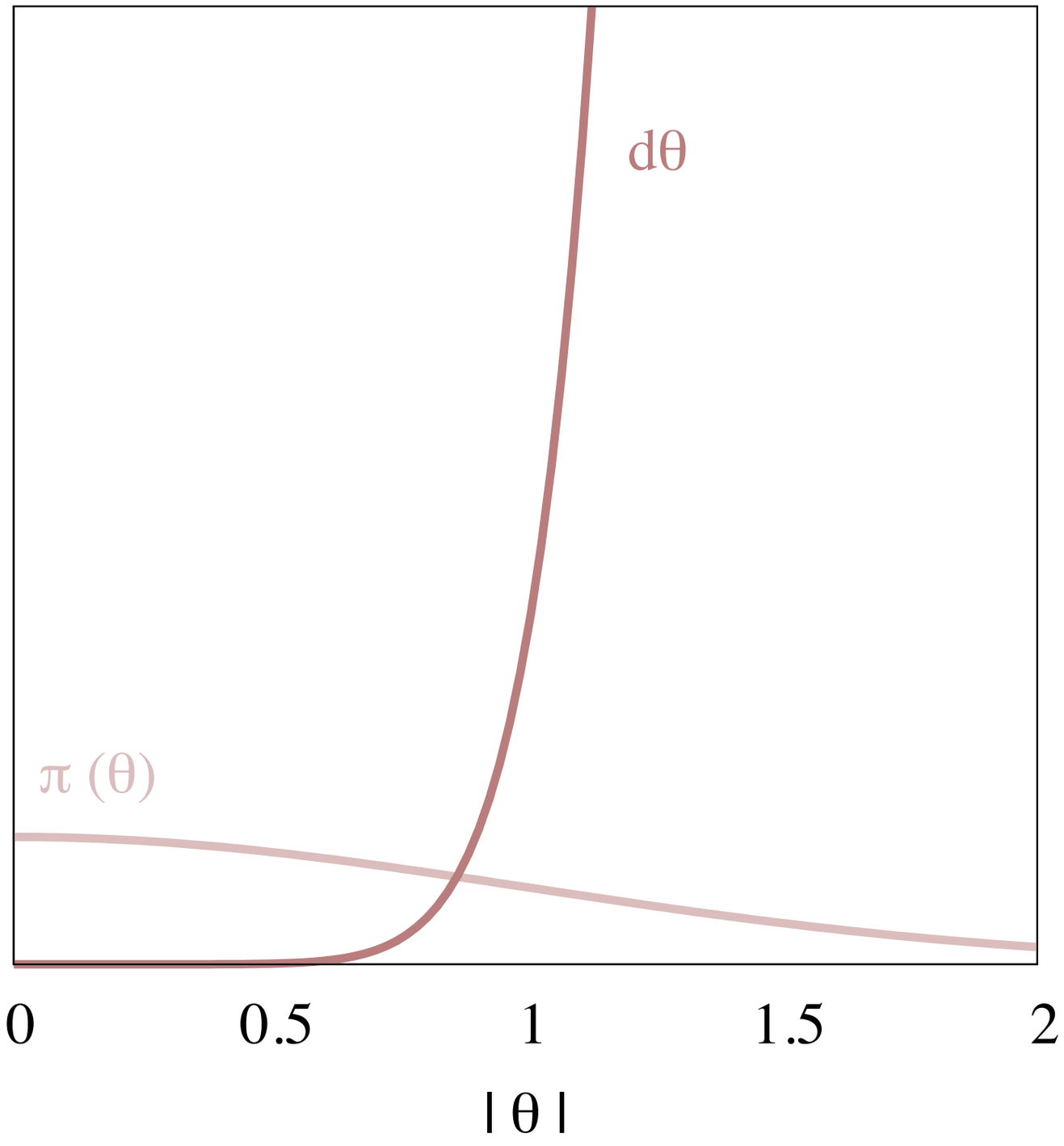
The ideal grid, however, is somewhat counterintuitive
-- it has to adapt to probability *mass*, not density.

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

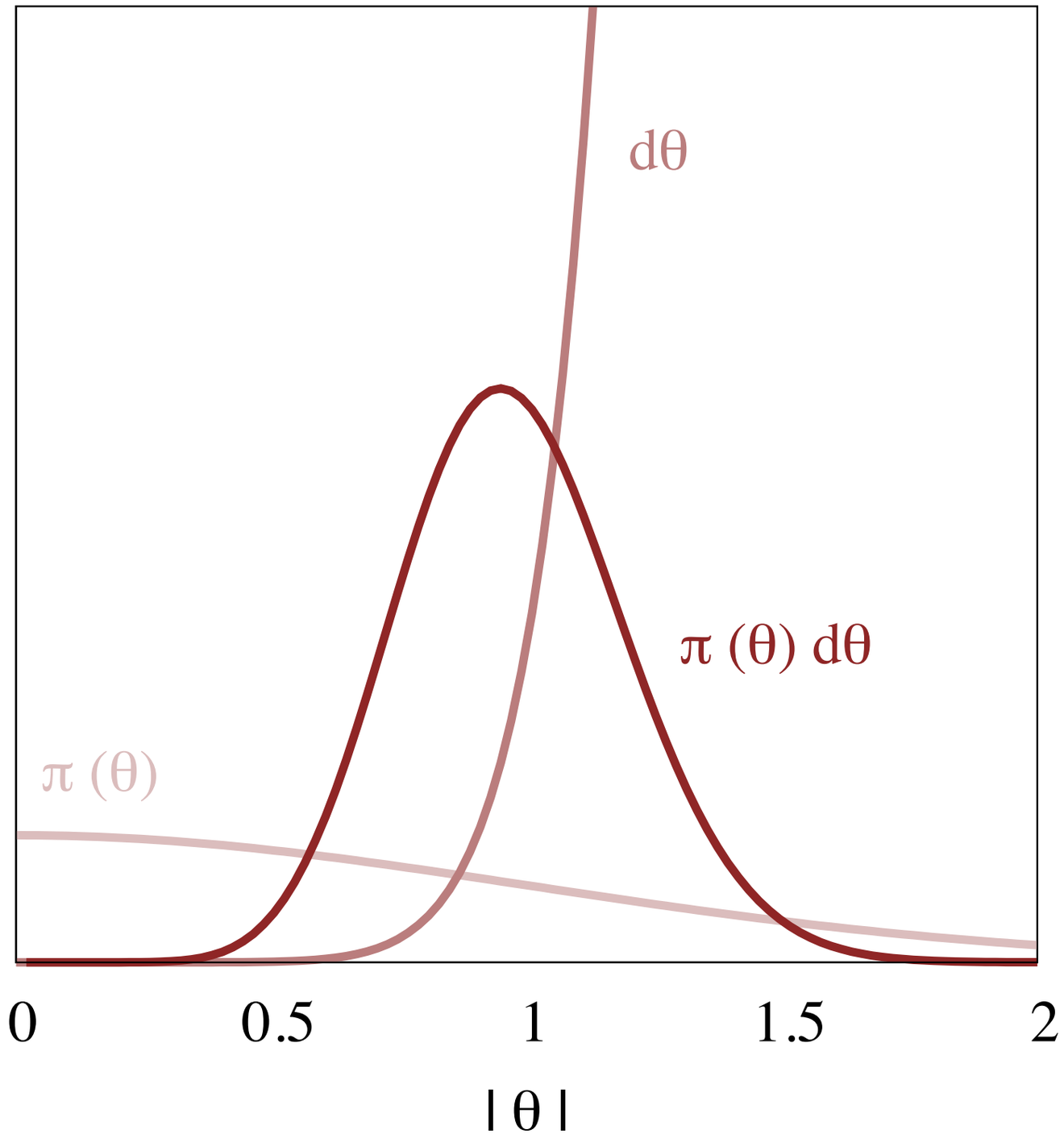
Because probability mass takes both density and volume into account it behaves very differently from either.



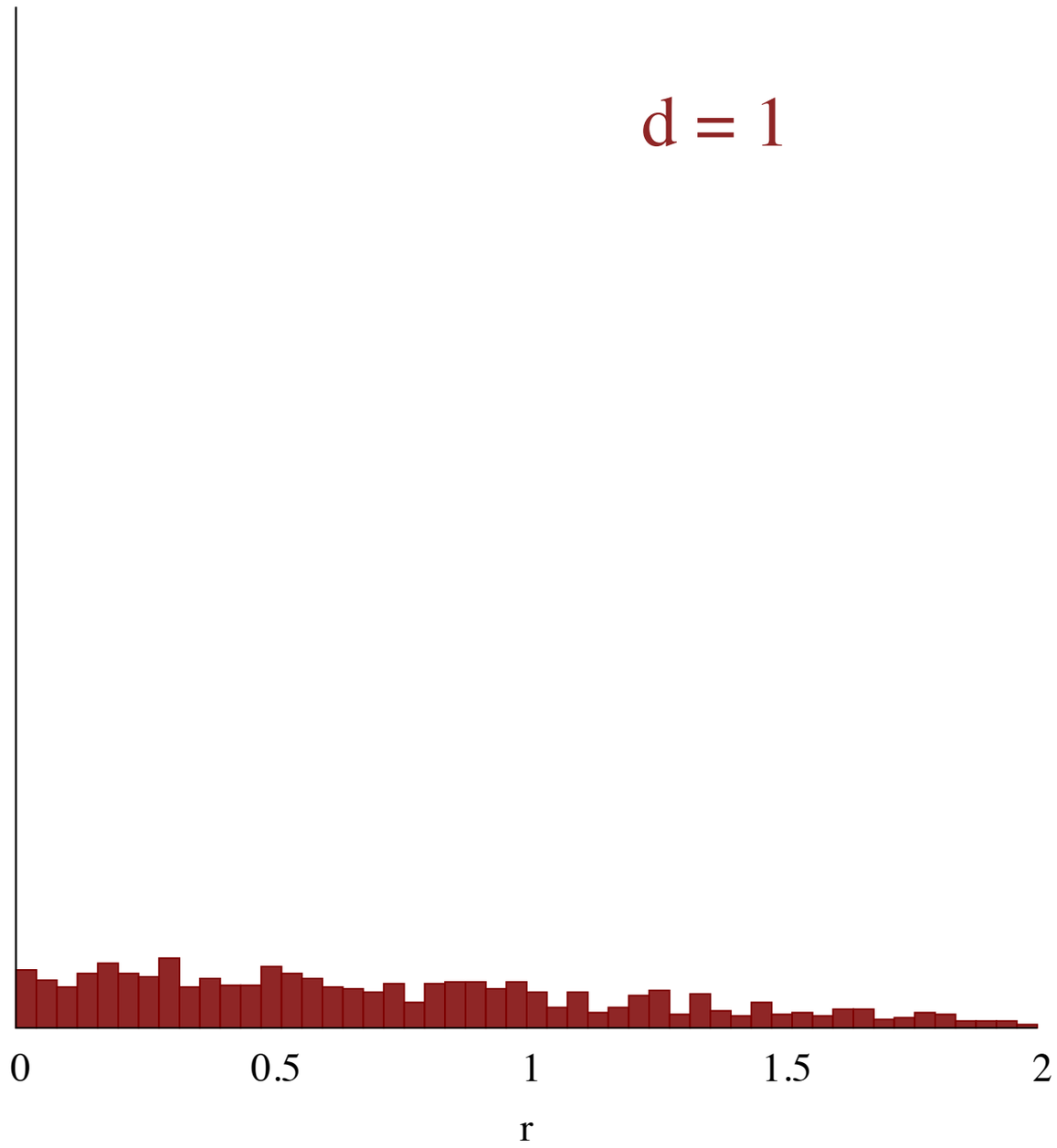
Because probability mass takes both density and volume into account it behaves very differently from either.



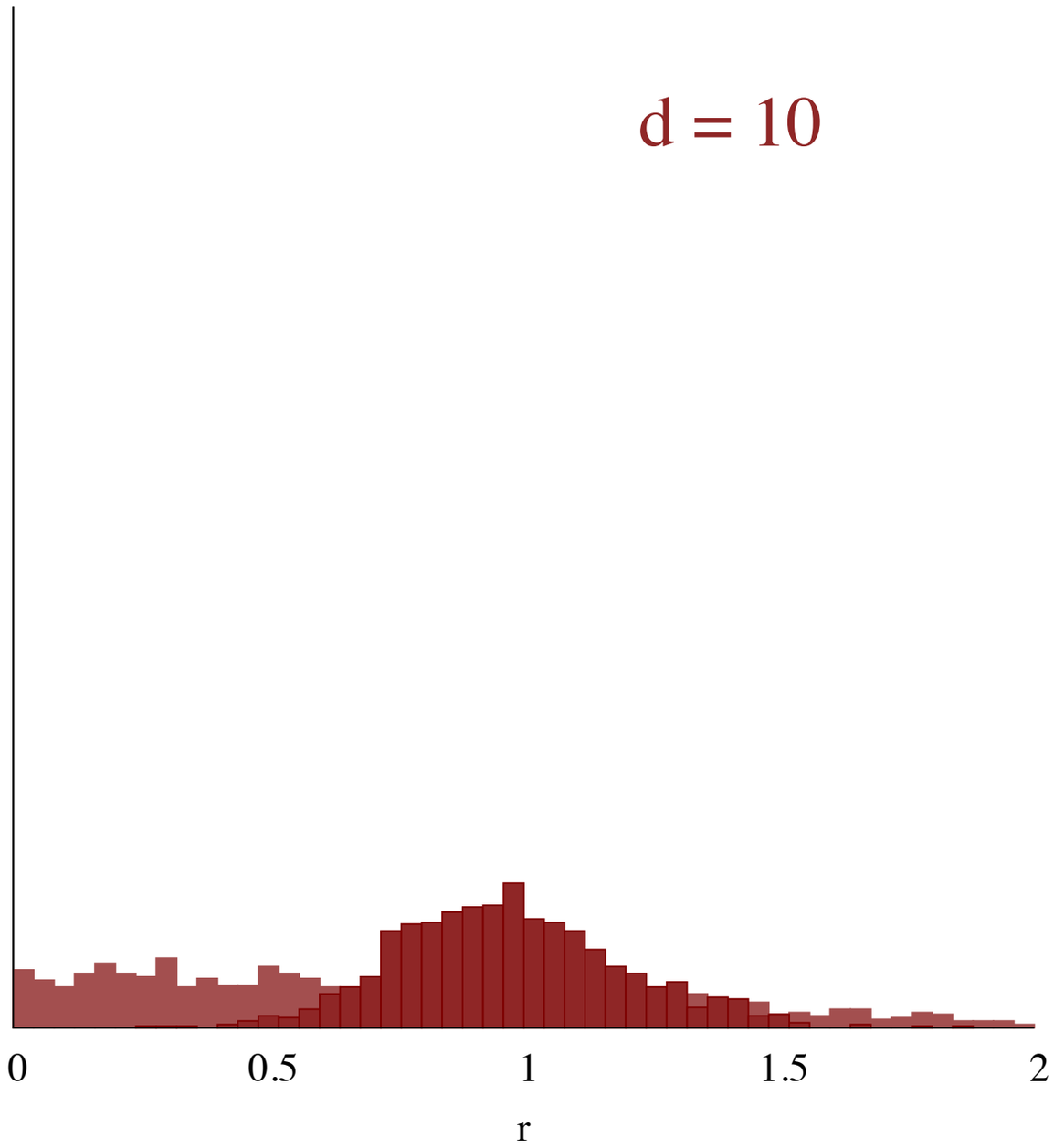
Because probability mass takes both density and volume into account it behaves very differently from either.



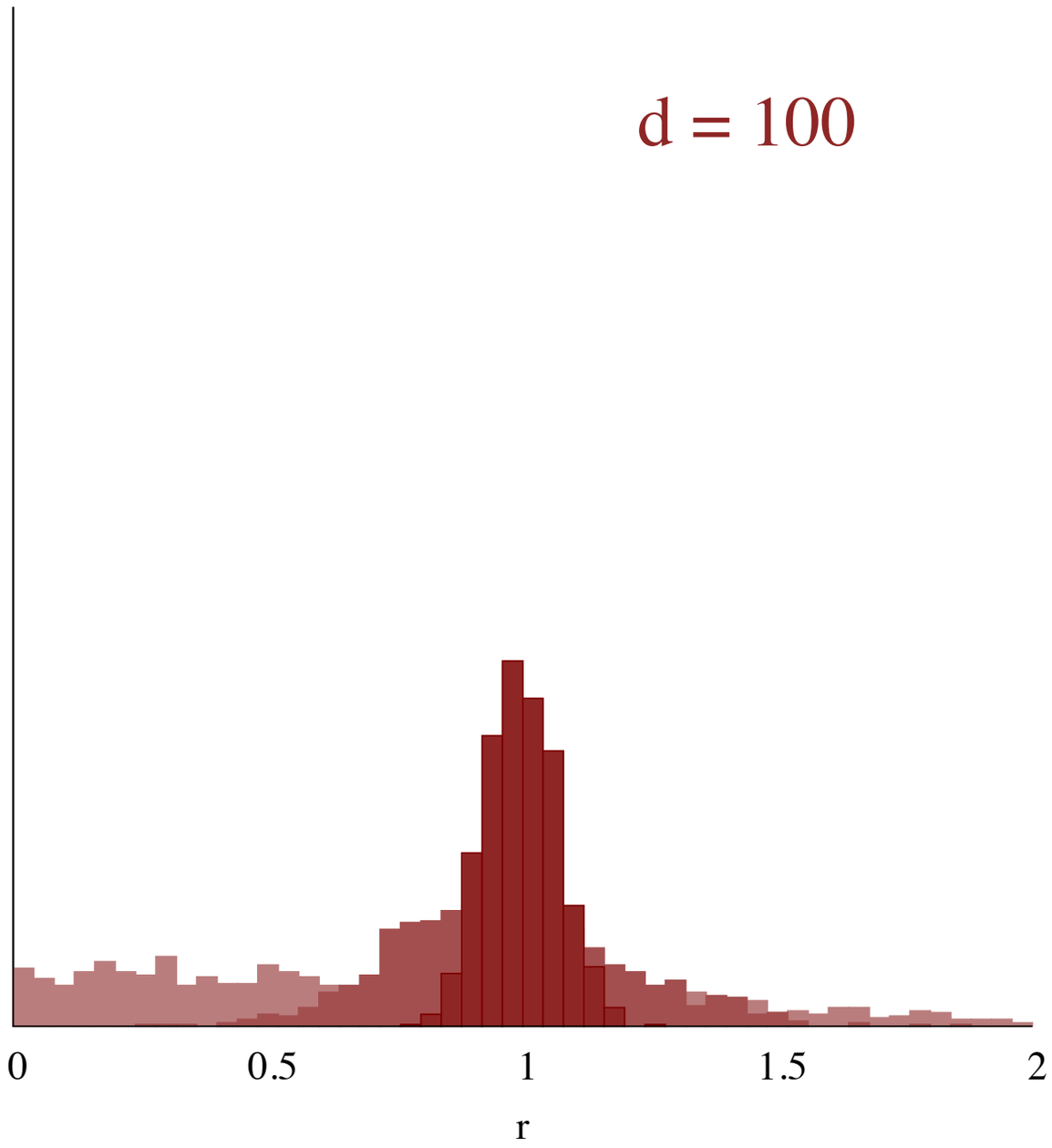
Instead probability mass concentrates on a hypersurface called the *typical set* that can be rather far from the mode.



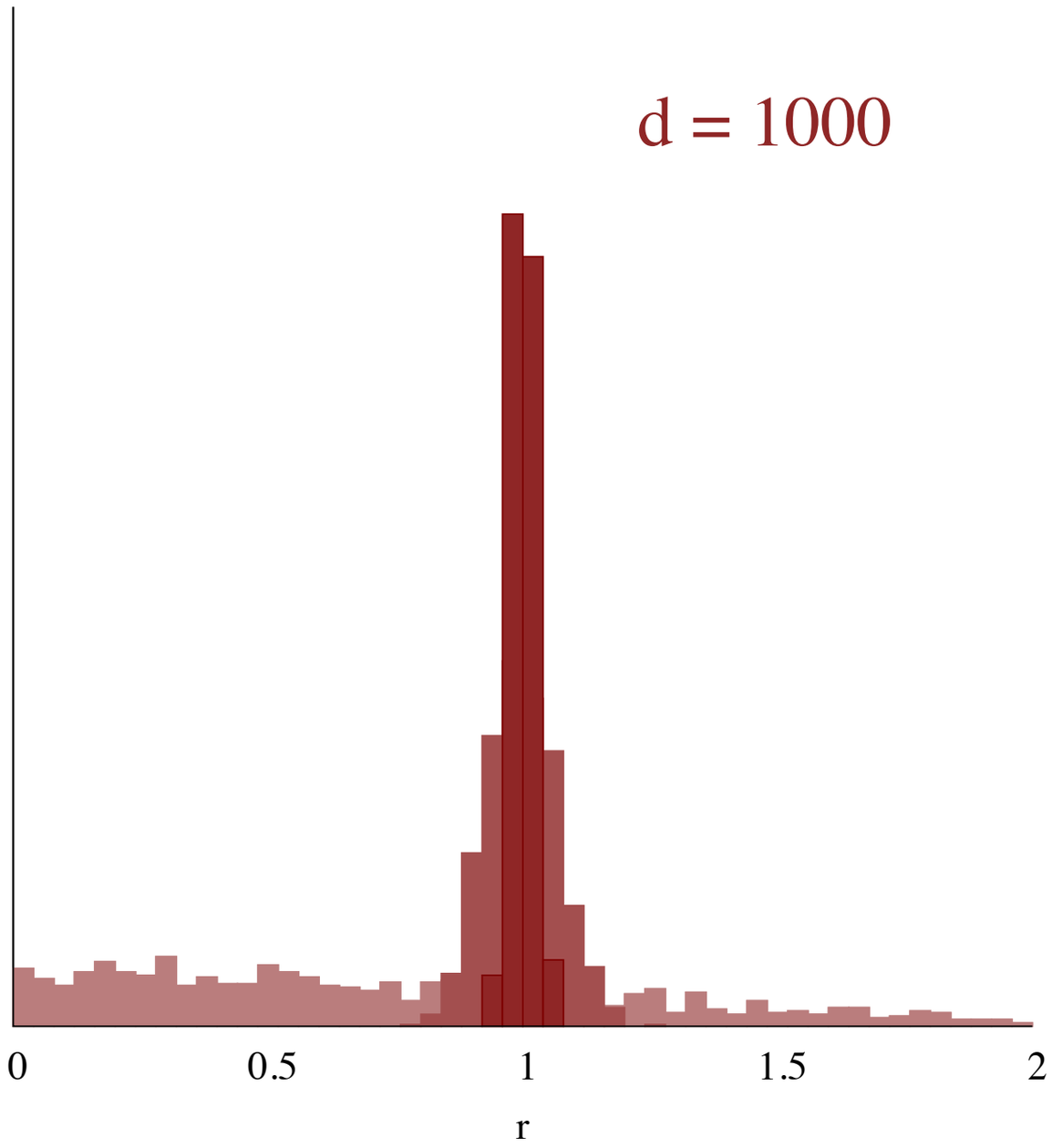
Instead probability mass concentrates on a hypersurface called the *typical set* that can be rather far from the mode.



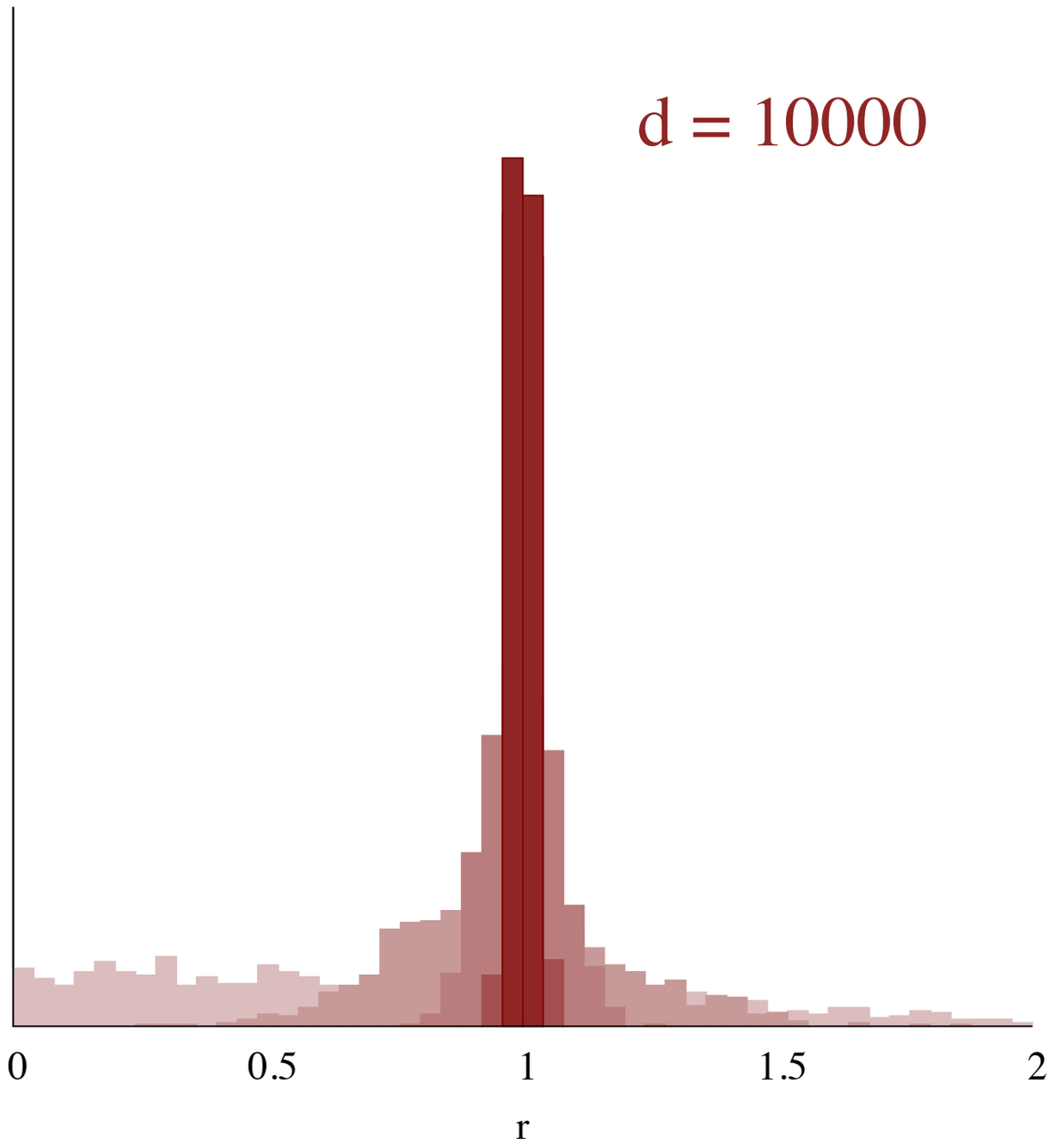
Instead probability mass concentrates on a hypersurface called the *typical set* that can be rather far from the mode.



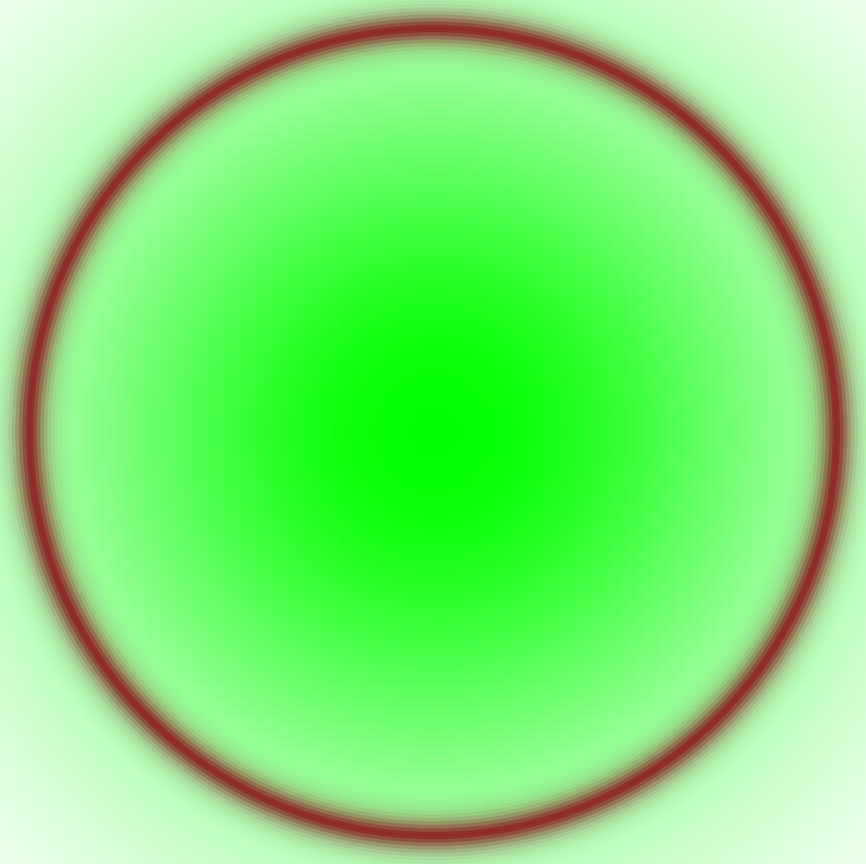
Instead probability mass concentrates on a hypersurface called the *typical set* that can be rather far from the mode.



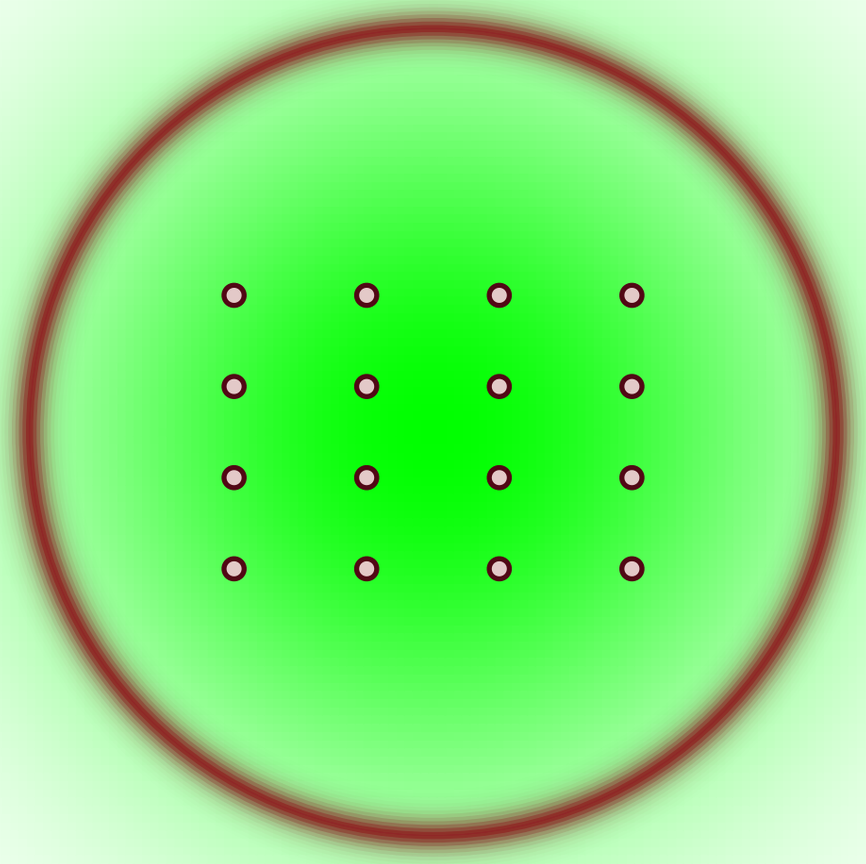
Instead probability mass concentrates on a hypersurface called the *typical set* that can be rather far from the mode.



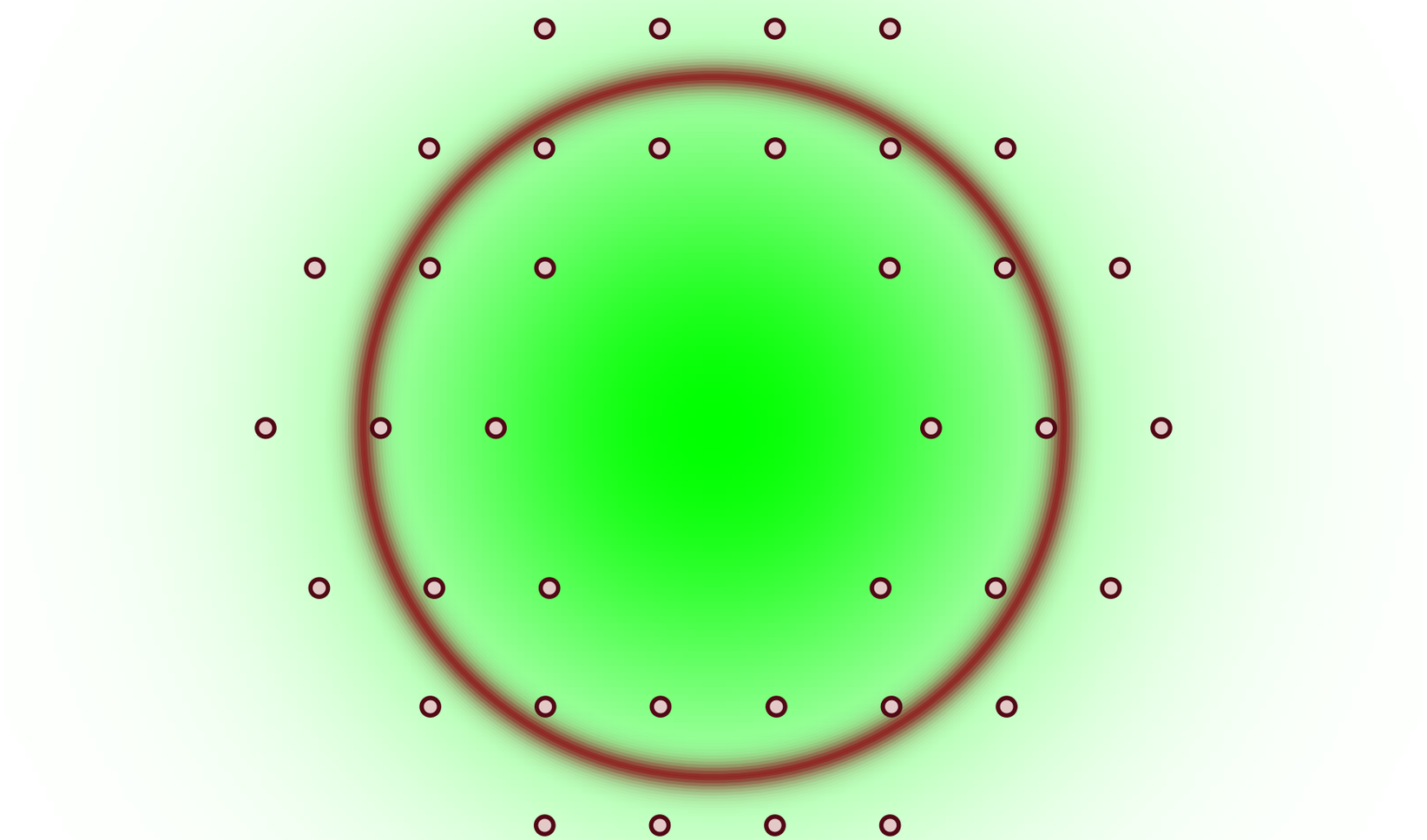
Constructing an efficient grid is intractable
even for seemingly simple distributions.



Constructing an efficient grid is intractable
even for seemingly simple distributions.



Constructing an efficient grid is intractable
even for seemingly simple distributions.



The problem becomes substantially more difficult for the complex distributions that arise in actual analyses.

Computational statistics is all about quantifying the typical set in order to approximating expectations.

Deterministic

Modal Estimators

Laplace Estimators

Variational Estimators

...

Stochastic

Rejection Sampling

Importance Sampling

Markov Chain Monte Carlo

...

Deterministic methods approximate the target expectations with those from a simpler distribution.

$$\pi(\theta) \approx q(\theta)$$

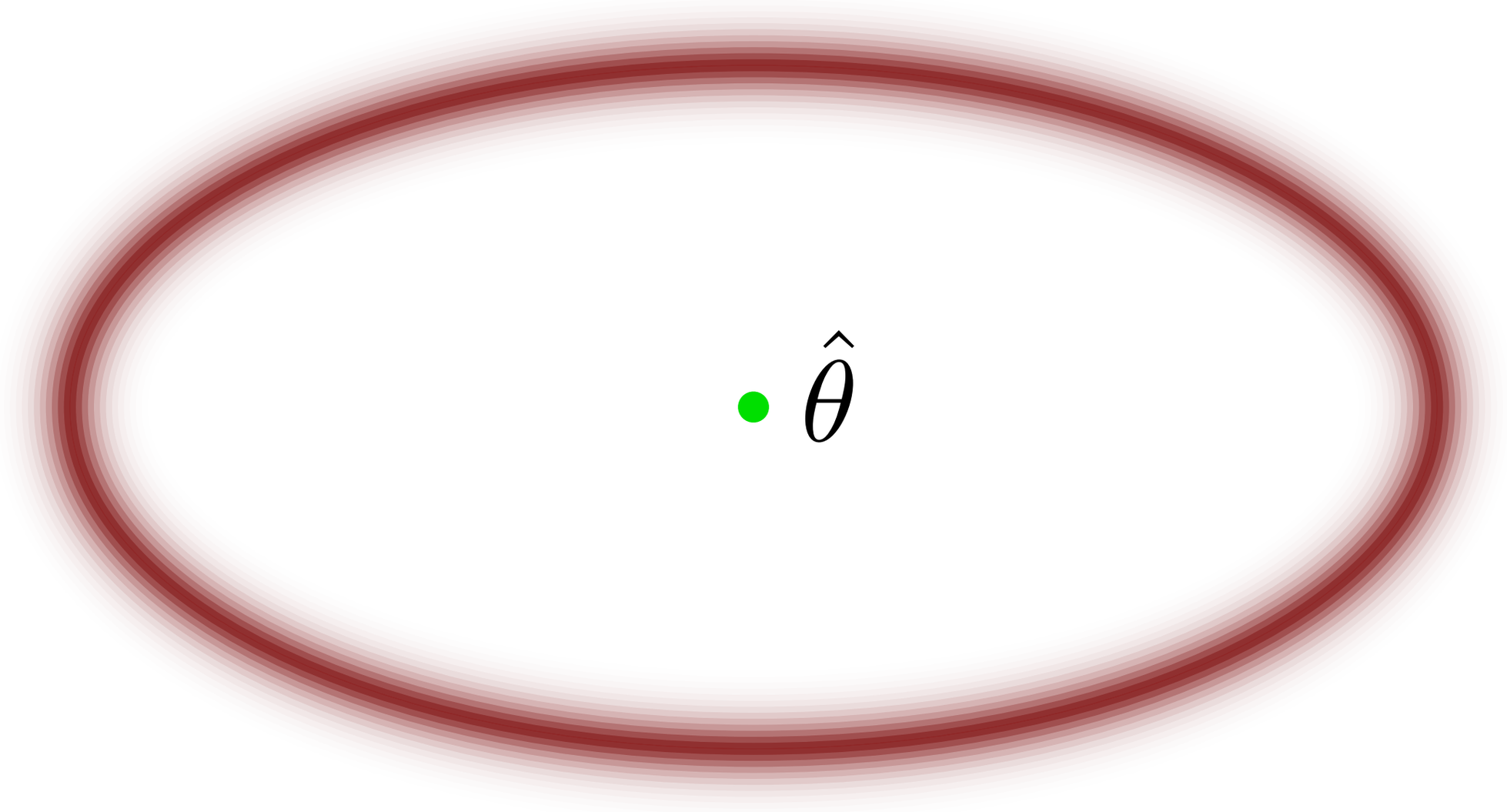
$$\int f(\theta)\, \pi(\theta)\, \mathrm{d}\theta \approx \int f(\theta)\, q(\theta)\, \mathrm{d}\theta$$

Modal estimators approximate expectations with point estimates at a density mode.

$$\pi(\theta) \approx \delta(\theta - \hat{\theta})$$

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

Given extreme symmetries, modal estimators of *some* functions in *some* parameterizations can be accurate.



Reparameterizations changes the density and hence the mode, but expectations should be invariant!

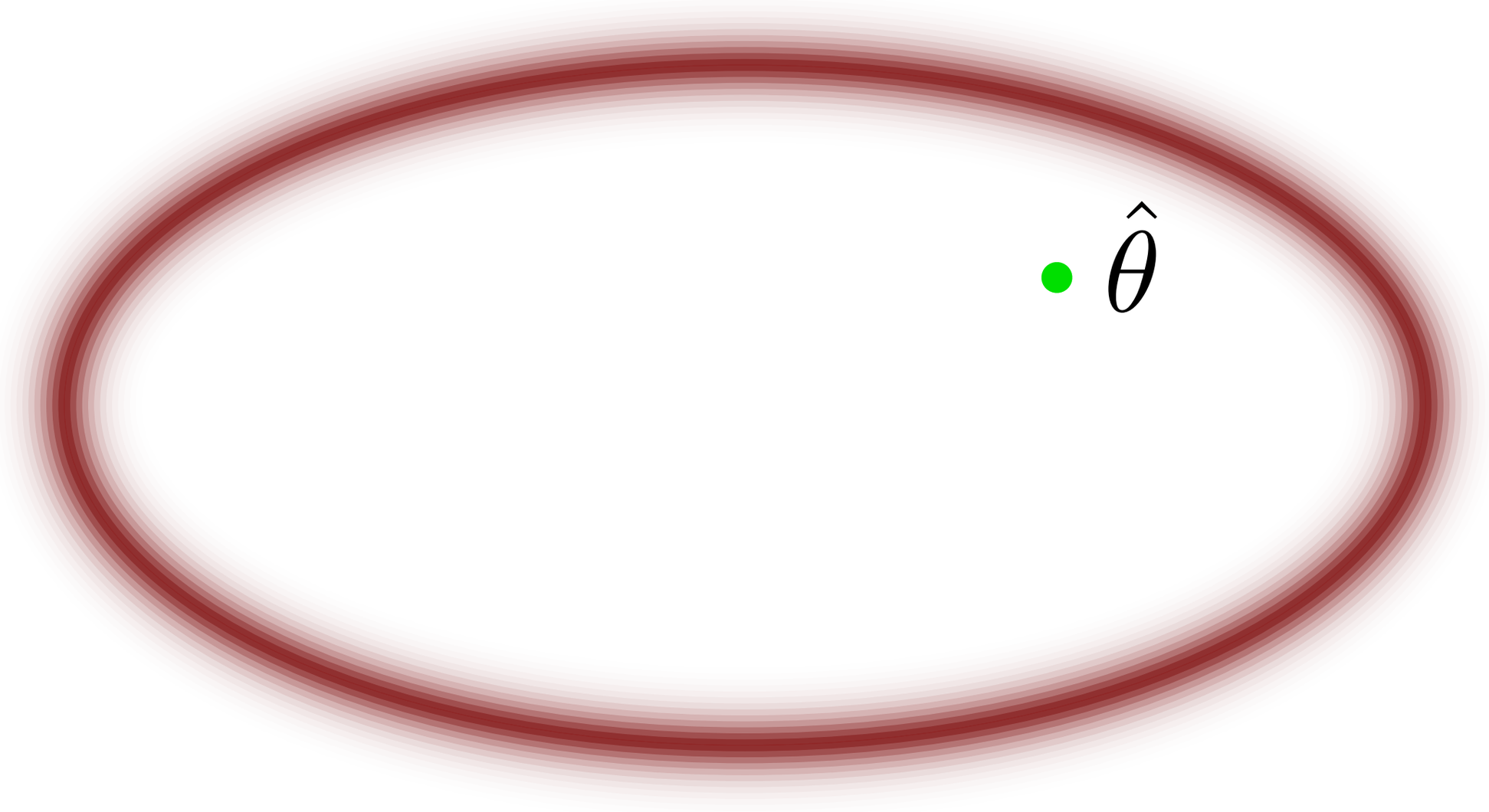
$$\mathbb{E}[f] = \int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta = \int f(\theta(\phi))\,\pi(\theta(\phi))\,\mathrm{d}\phi$$

Reparameterizations changes the density and hence the mode, but expectations should be invariant!

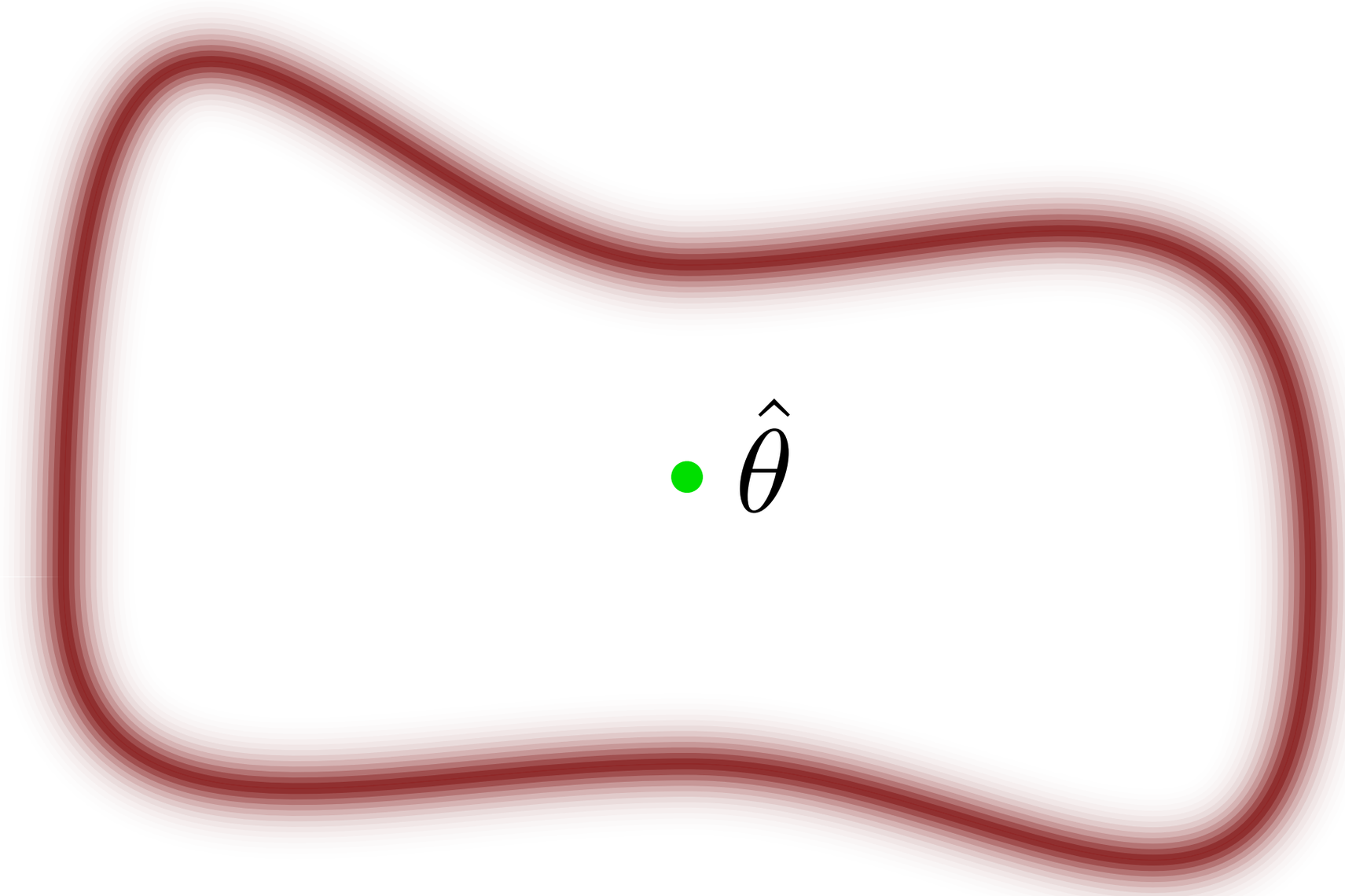
$$\mathbb{E}[f] = \int f(\theta)\,\pi(\theta)\,\mathrm{d}\theta = \int f(\theta(\phi))\,\pi(\theta(\phi))\,\mathrm{d}\phi$$

$$\pi(\theta(\phi)) = \pi(\theta)\left|\frac{\partial\theta}{\partial\phi}\right|$$

Consequently, any reparameterization will drastically affect the performance of a modal estimators.



For the complex typical sets in realistic problems, there is likely no parameterization that works well.

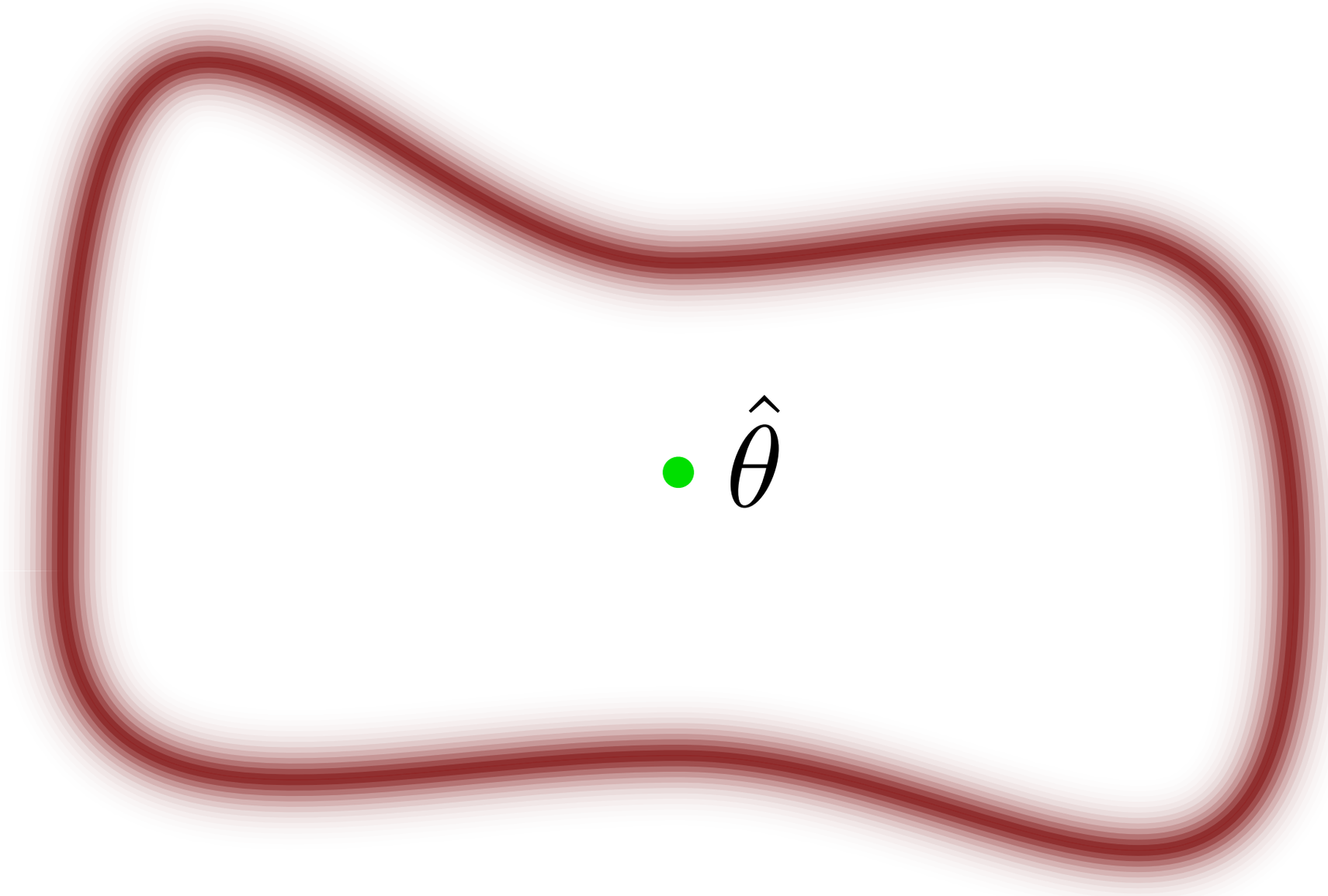


Laplace approximations complement modal estimators with a Hessian to quantify the breadth of the typical set.

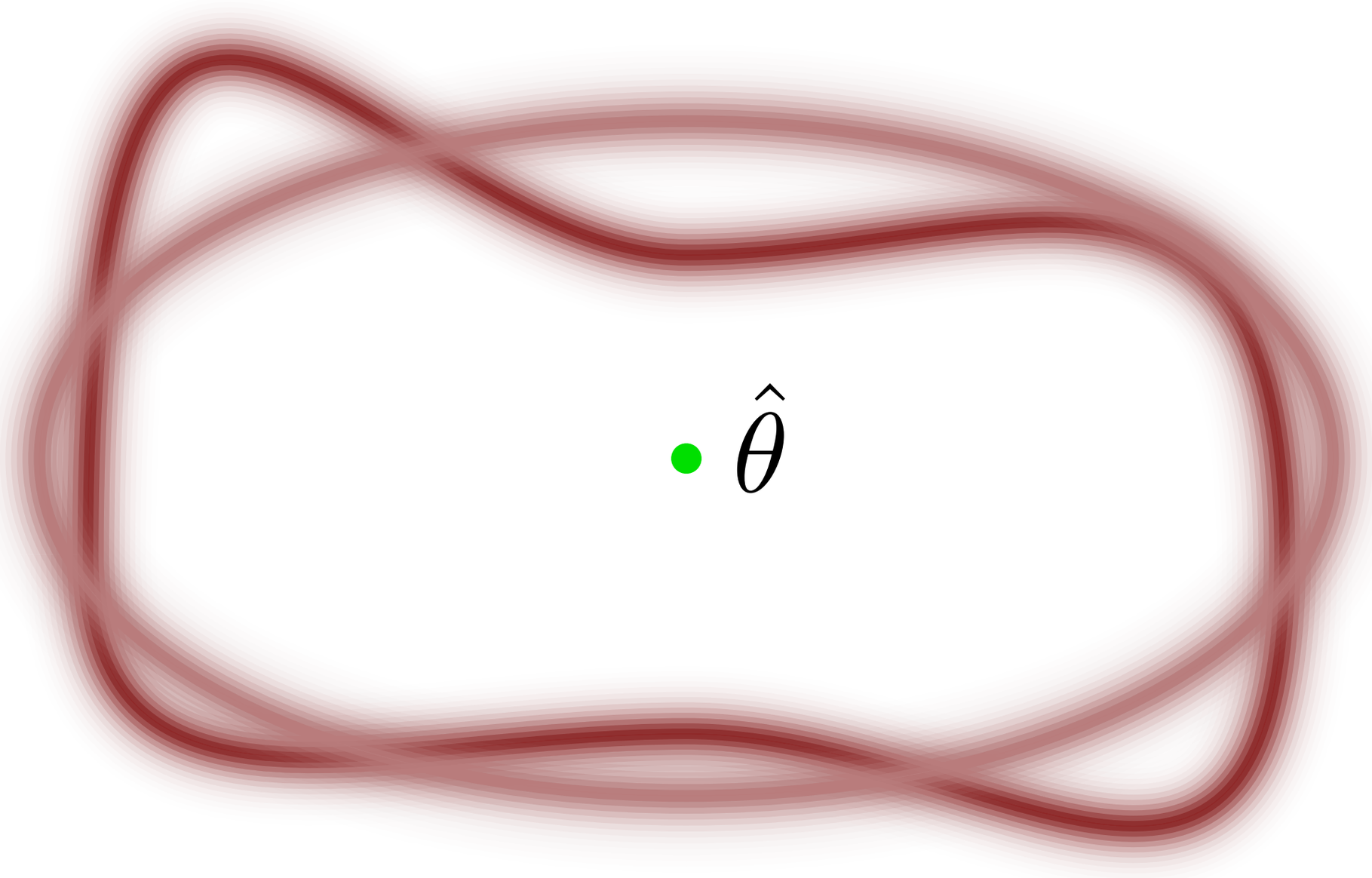
$$\pi(\theta) \approx \mathcal{N}(\theta \mid \hat{\theta}, H^{-1})$$

$$H = \left. \frac{\partial^2 \pi(\theta)}{\partial \theta^2} \right|_{\theta = \hat{\theta}}$$

Laplace approximations complement modal estimators with a Hessian to quantify the breadth of the typical set.



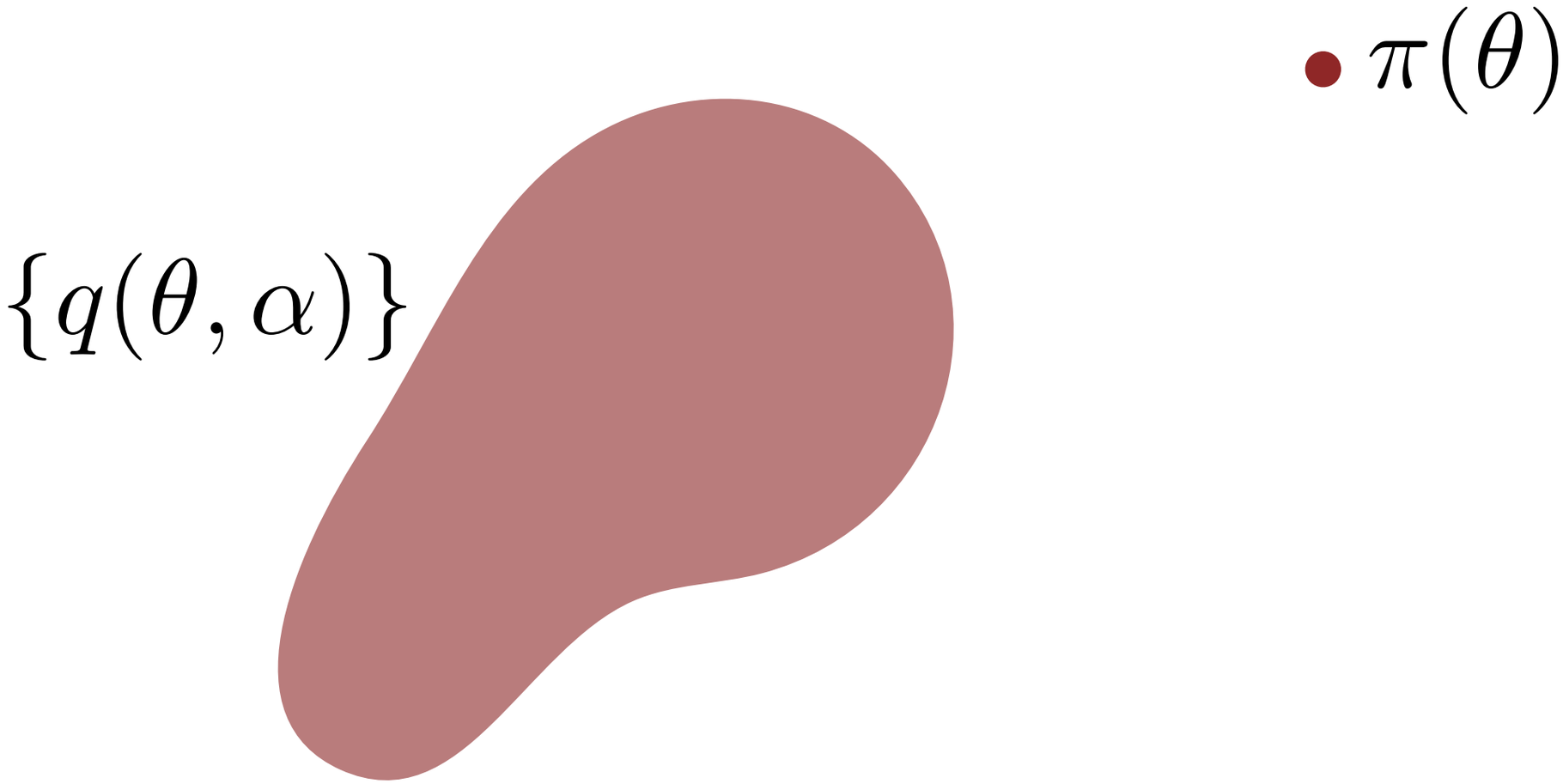
Laplace approximations complement modal estimators with a Hessian to quantify the breadth of the typical set.



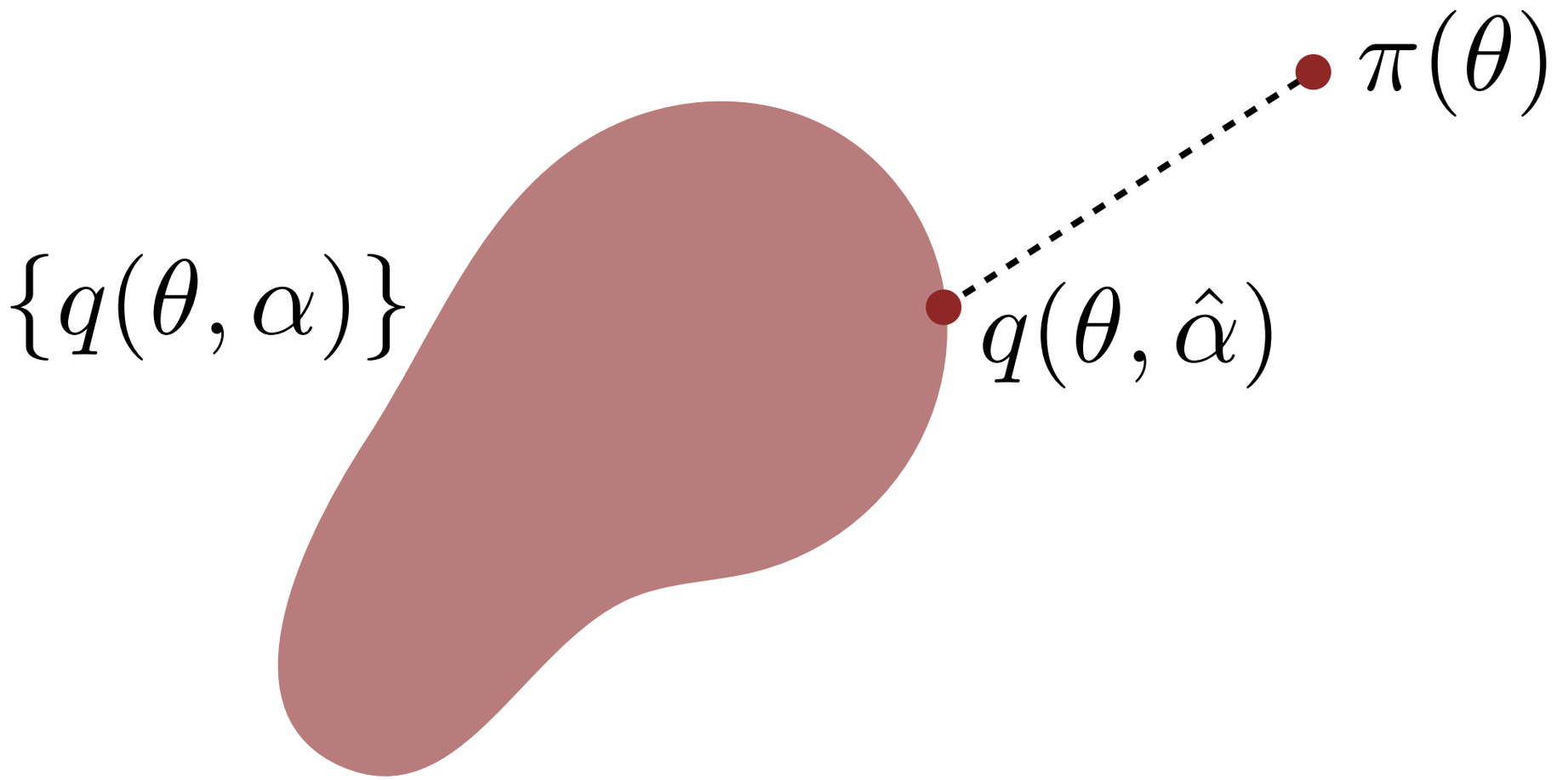
Variational methods turn the approximation problem into a variational optimization.

- $\pi(\theta)$

Variational methods turn the approximation problem into a variational optimization.



Variational methods turn the approximation problem into a variational optimization.



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

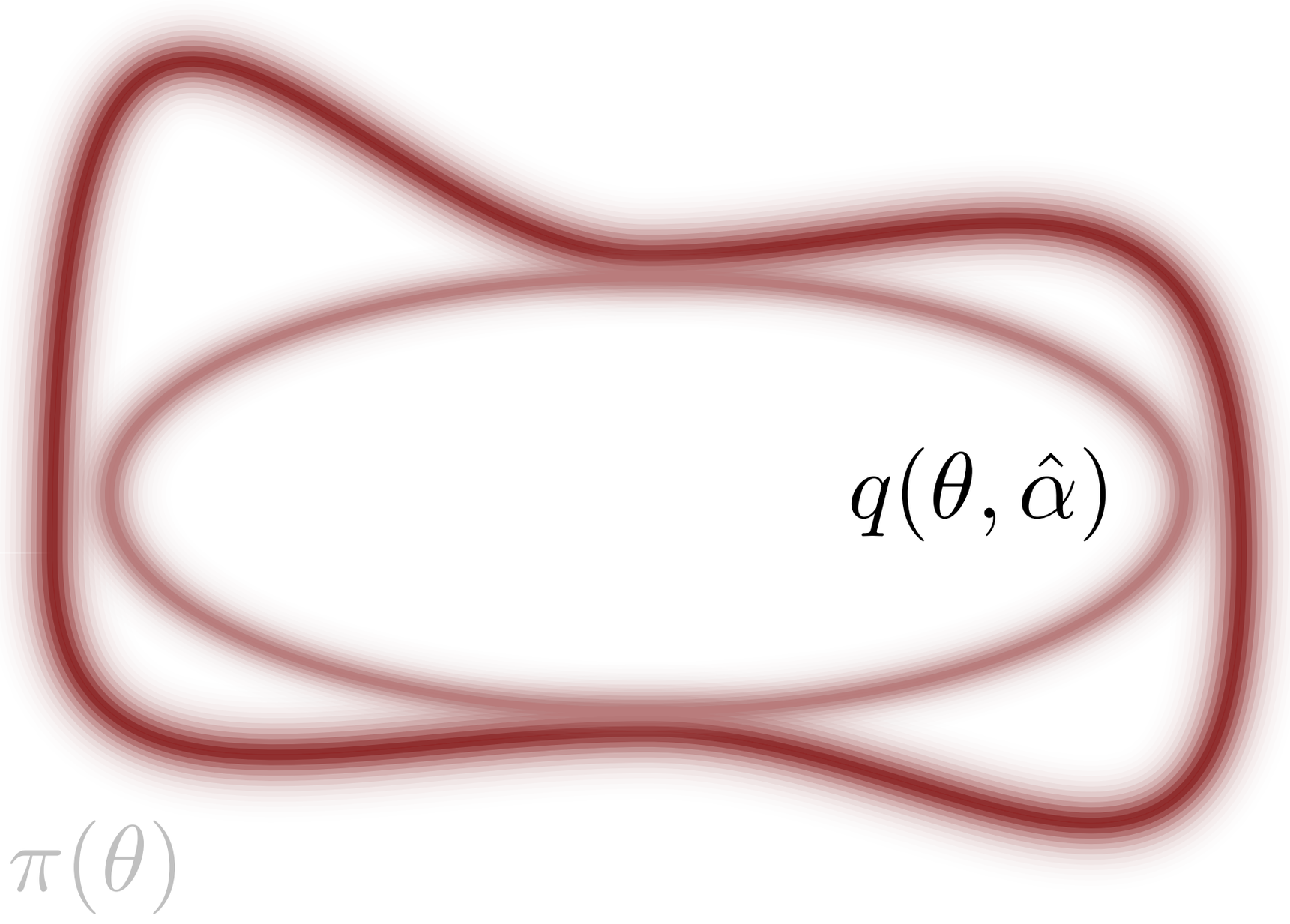
$$\pi(\theta) \approx q(\theta, \hat{\alpha})$$

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

And then use the closet element of the variational family to approximate posterior expectations.

$\pi(\theta)$

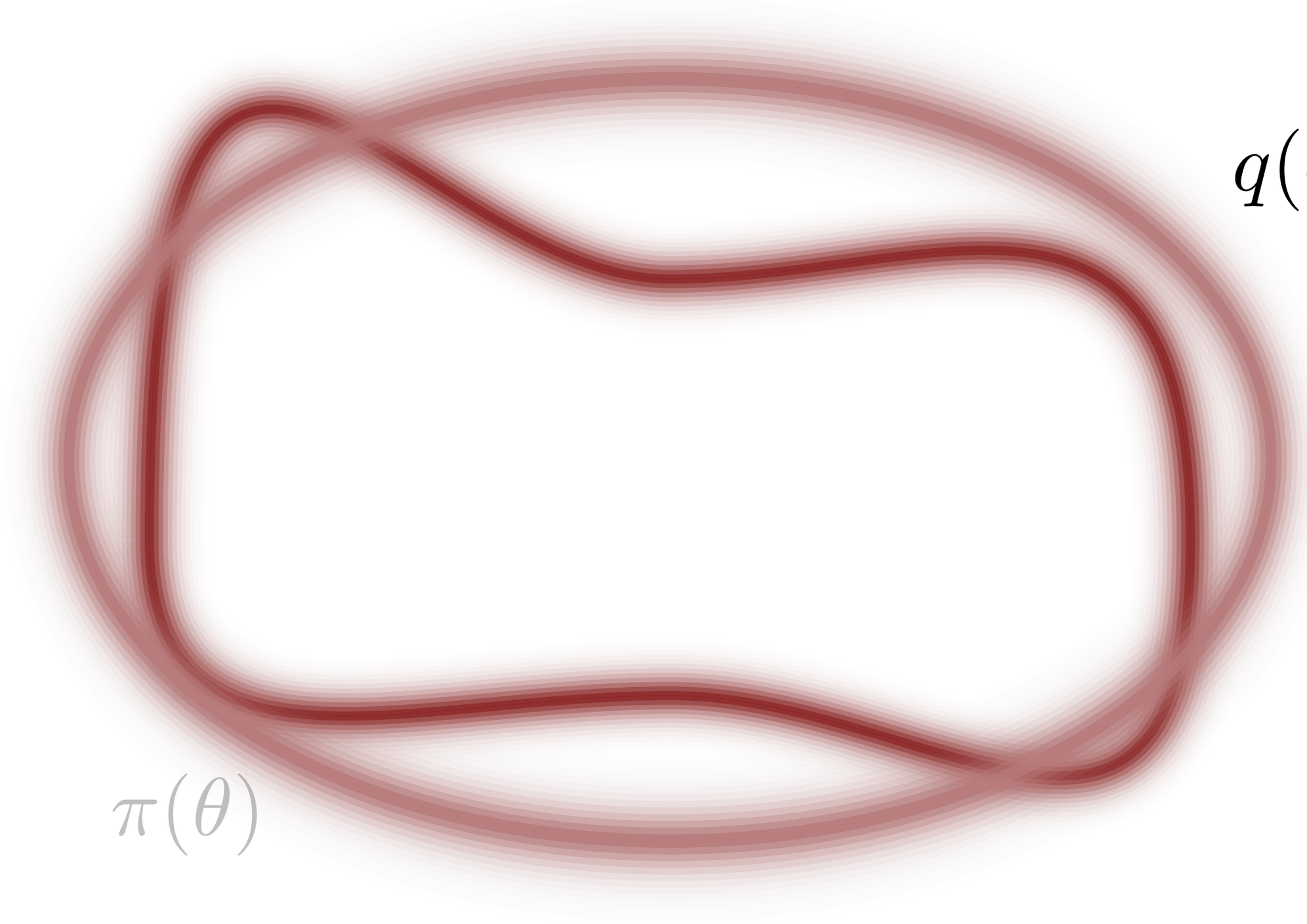
And then use the closet element of the variational family to approximate posterior expectations.



And then use the closet element of the variational family to approximate posterior expectations.

$\pi(\theta)$

And then use the closet element of the variational family to approximate posterior expectations.



Stochastic methods construct random estimators of the exact expectations.

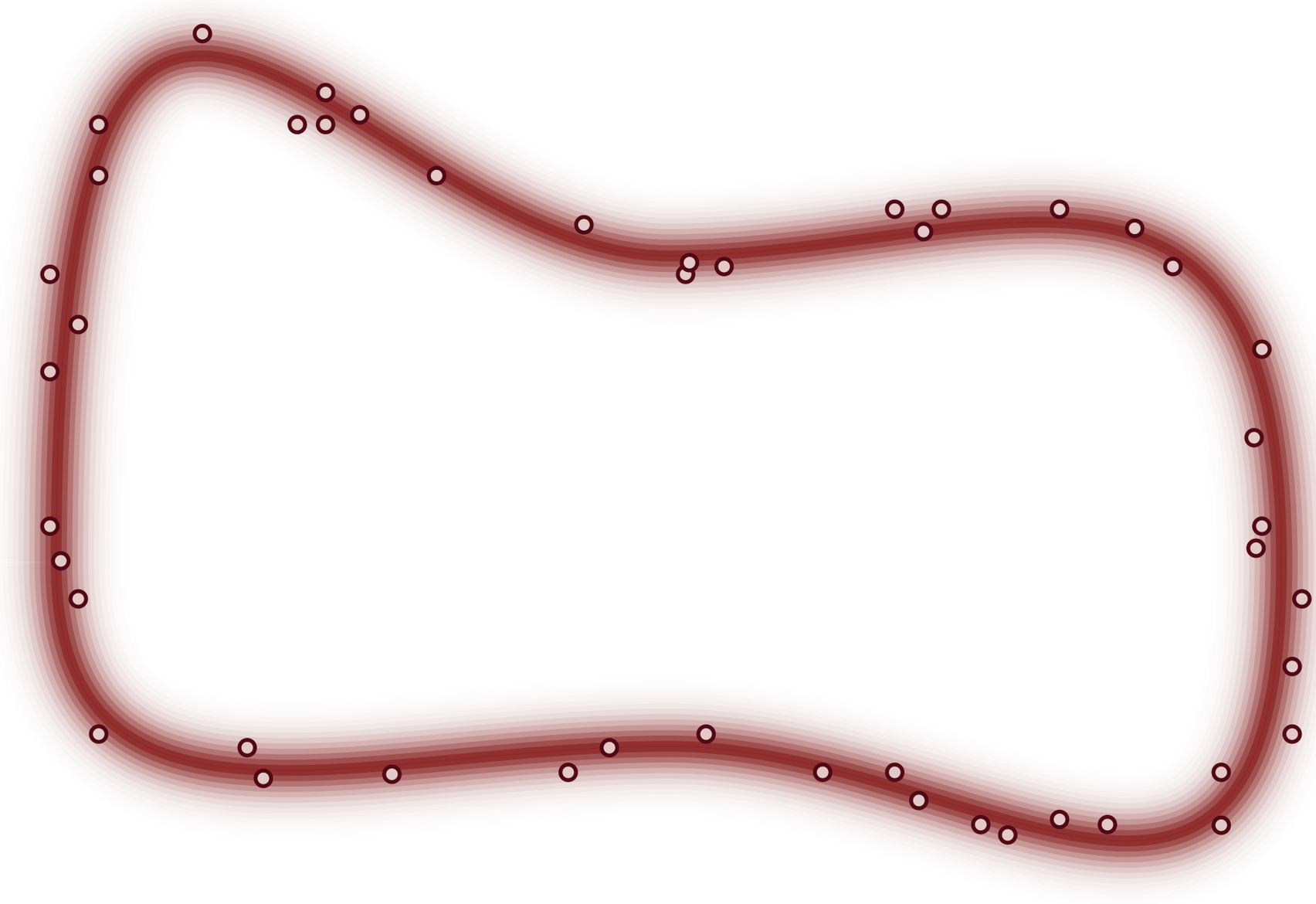
$$\{\theta_1, \ldots, \theta_N\}$$

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

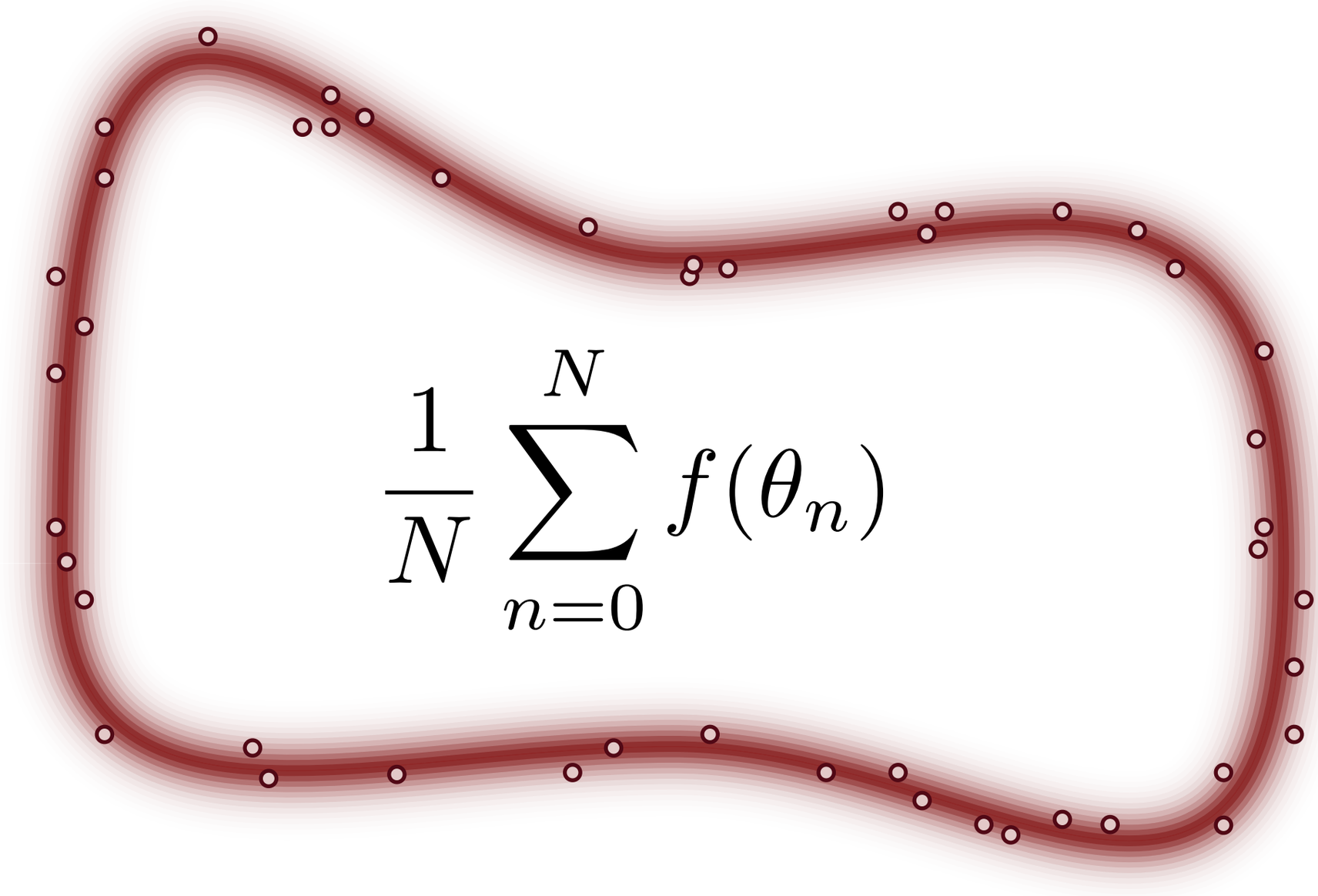
The foremost stochastic techniques are *Monte Carlo* methods.

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

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



Monte Carlo estimators average the functions over these samples, eventually converging to the true expectation.



Monte Carlo estimators average the functions over these samples, eventually converging to the true expectation.

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

In practice we can't generate exact samples from complex distributions, but we can generate *correlated* samples.

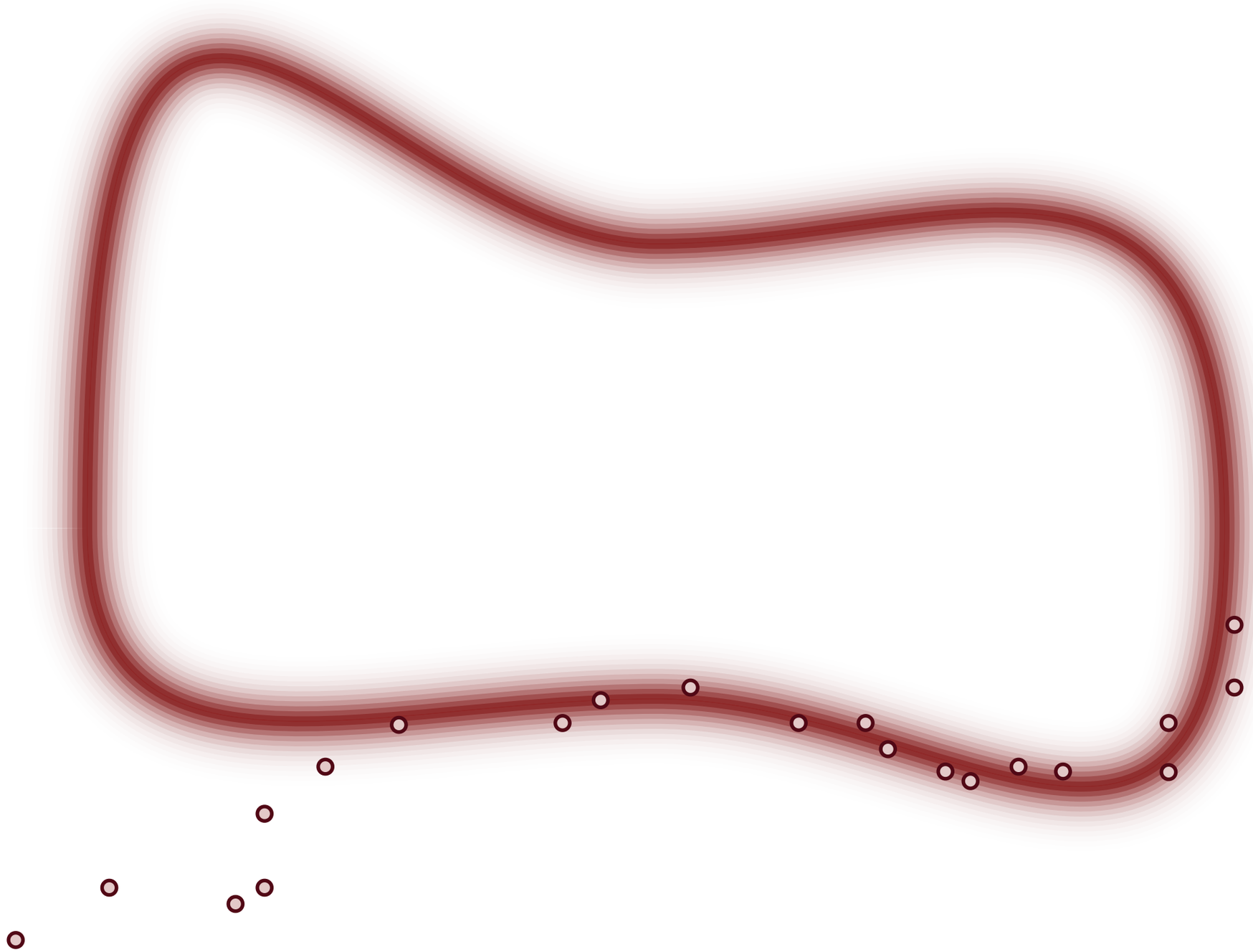
$$T(\theta \mid \theta')$$

In practice we can't generate exact samples from complex distributions, but we can generate *correlated* samples.

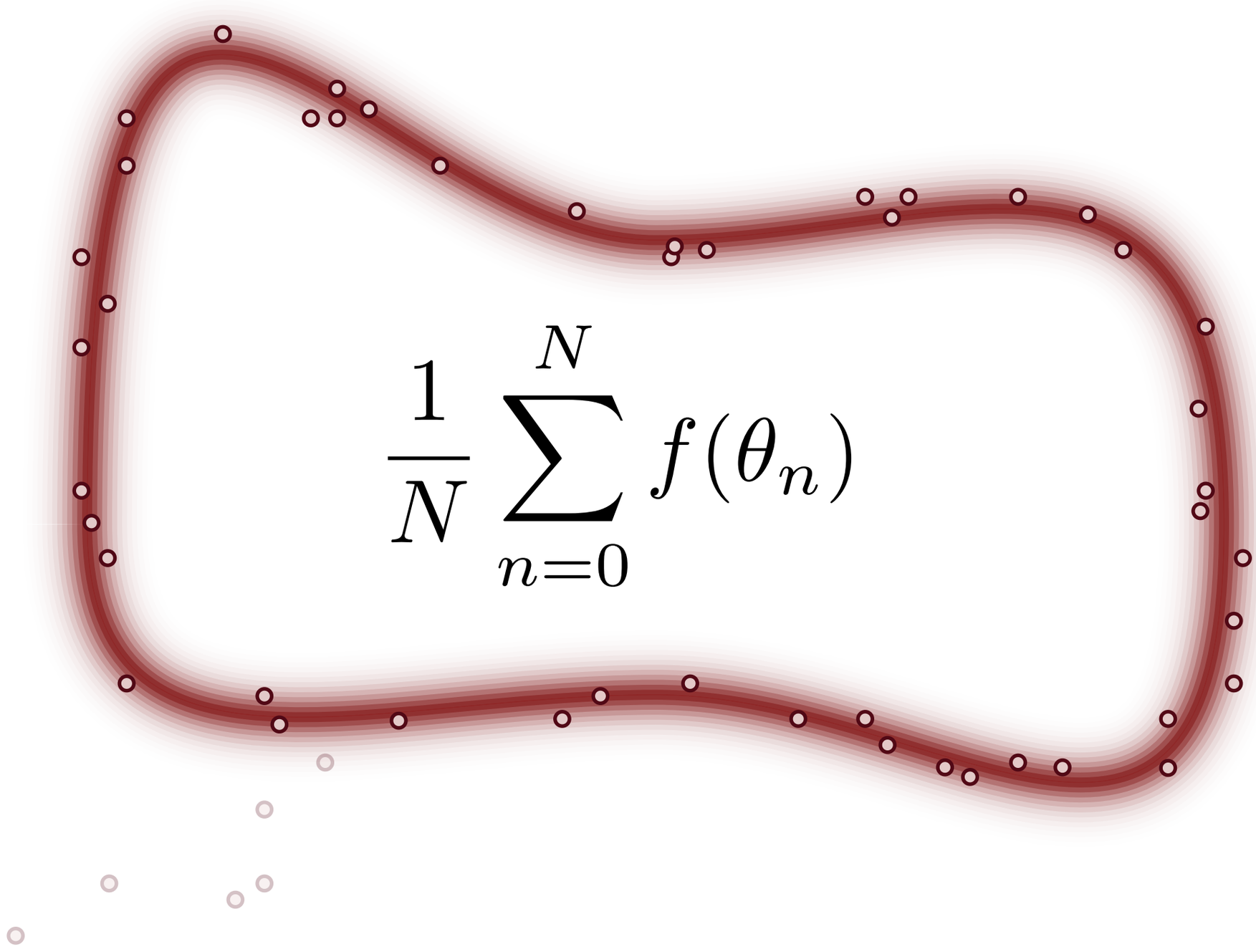
$$\pi(\theta) = \int T(\theta \mid \theta') \, \pi(\theta') \, \mathrm{d}\theta'$$

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

Markov chains 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*.

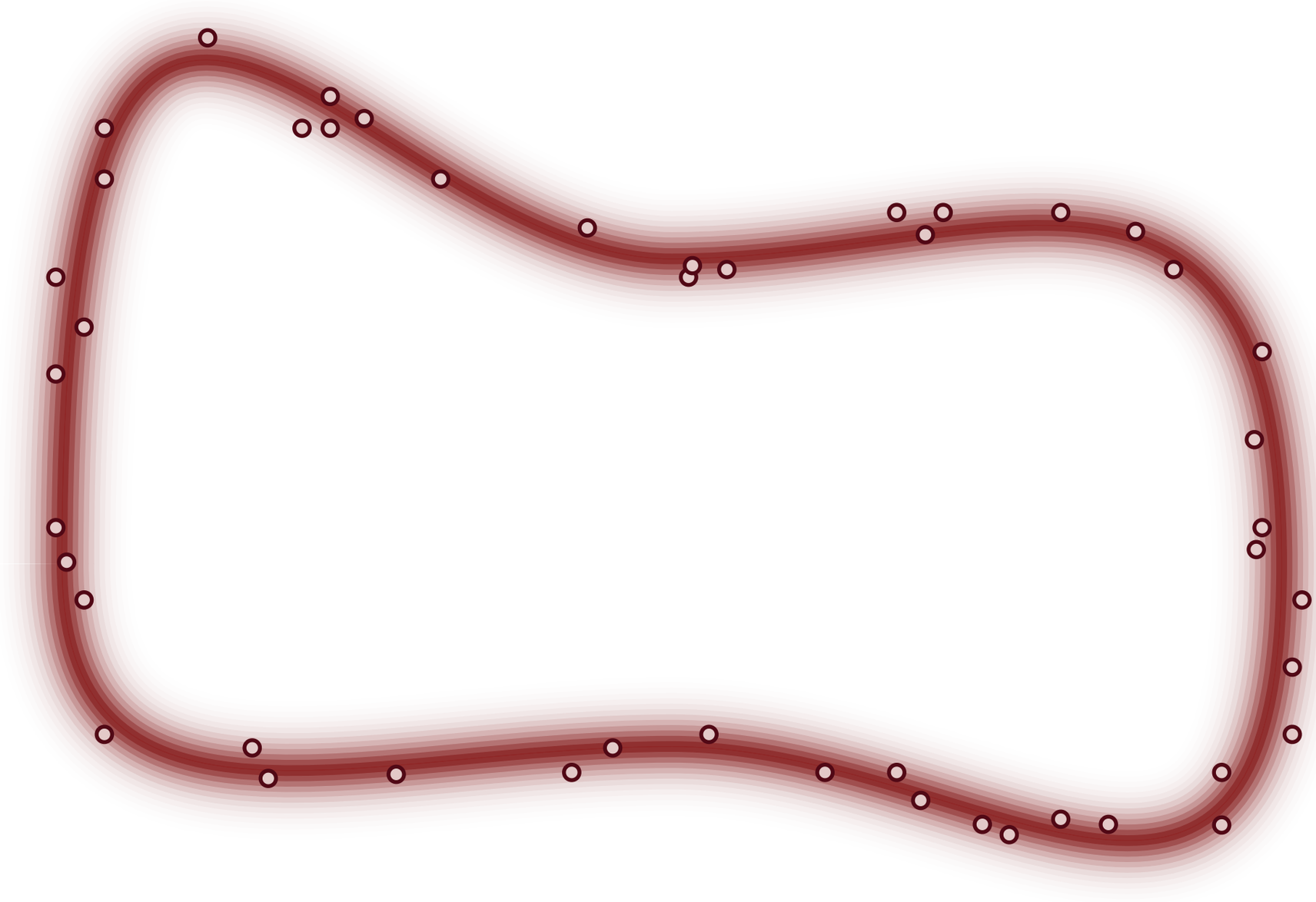


$$\frac{1}{N} \sum_{n=0}^{N} f(\theta_n)$$

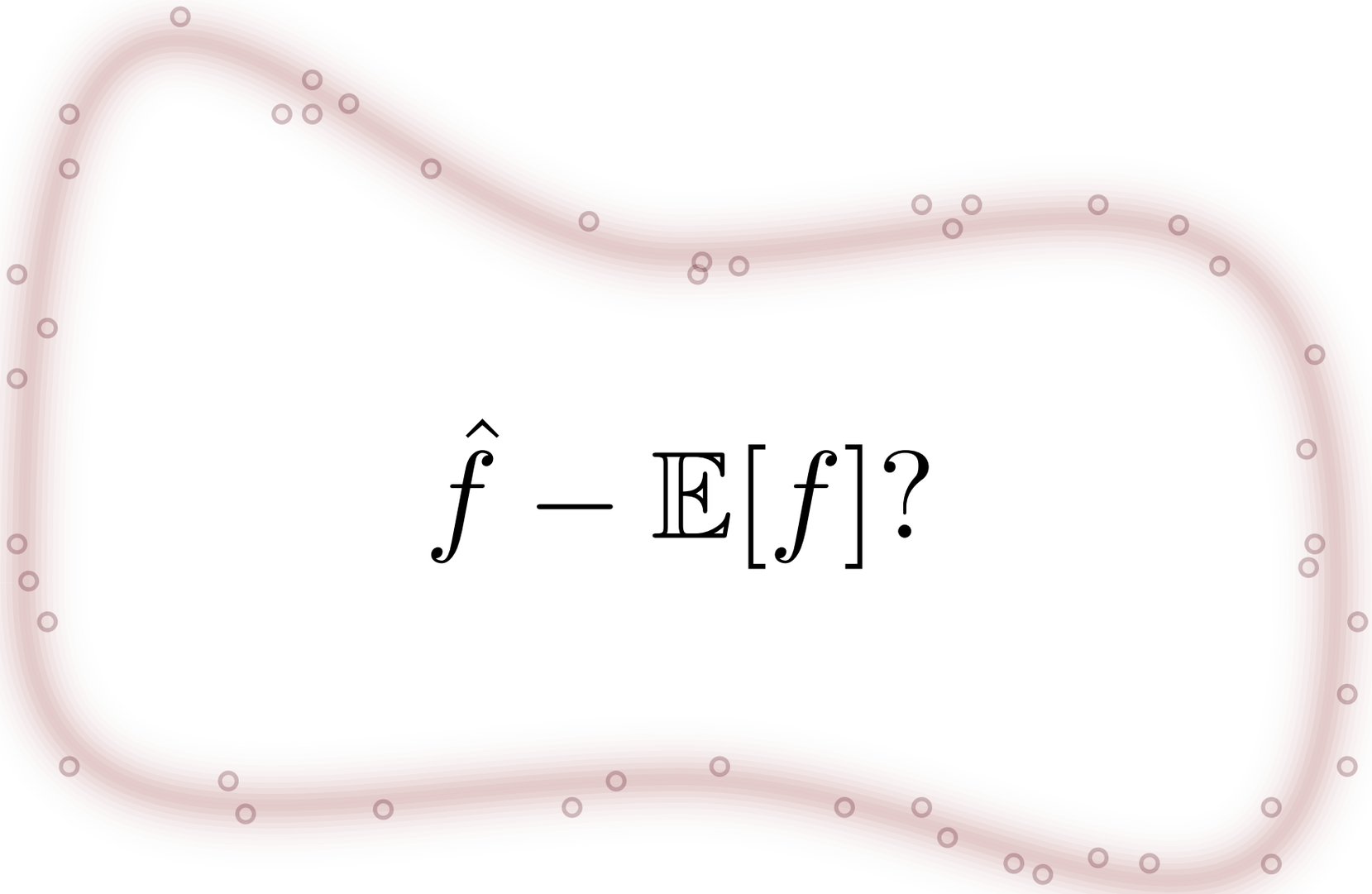
If run long the chain long enough then we can construct consistent *Markov Chain Monte Carlo estimators*.

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

Ultimately, there are many of methods for quantifying the typical set and, consequently, integrals.



To use any method responsibly, however, you have to be able to quantify the accuracy of the estimation!



$$\hat{f} - \mathbb{E}[f]?$$

https://github.com/betanalpha/stan_intro