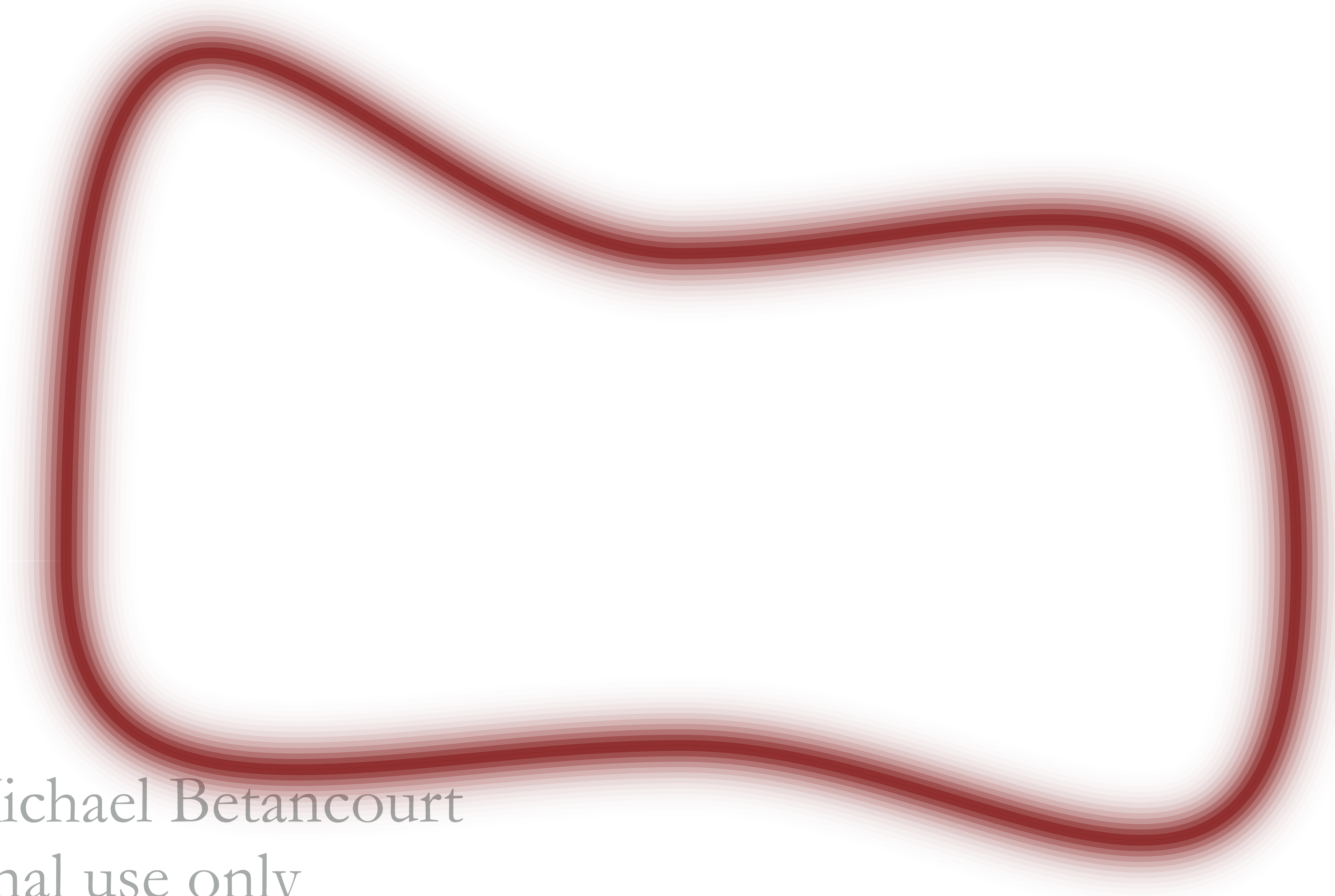
Concentration of Measure



Michael Betancourt @betanalpha
Symplectomorphic, LLC

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

In the Bayesian paradigm inferences are encoding in a probability distribution given by *Bayes' Theorem*.

$$\pi_S(\theta \mid \tilde{y}) = \frac{\pi_S(\tilde{y} \mid \theta)}{\pi_S(\tilde{y})} \pi_S(\theta)$$

The *prior distribution* quantifies relevant domain expertise available before a measurement is made.

$$\pi_S(\theta \mid \tilde{y}) = \frac{\pi_S(\tilde{y} \mid \theta)}{\pi_S(\tilde{y})} \pi_S(\theta)$$

The *likelihood function* quantifies what we learn about the model configurations from any given measurement.

$$\pi_S(\theta \mid \tilde{y}) = \frac{\pi_S(\tilde{y} \mid \theta)}{\pi_S(\tilde{y})} \pi_S(\theta)$$

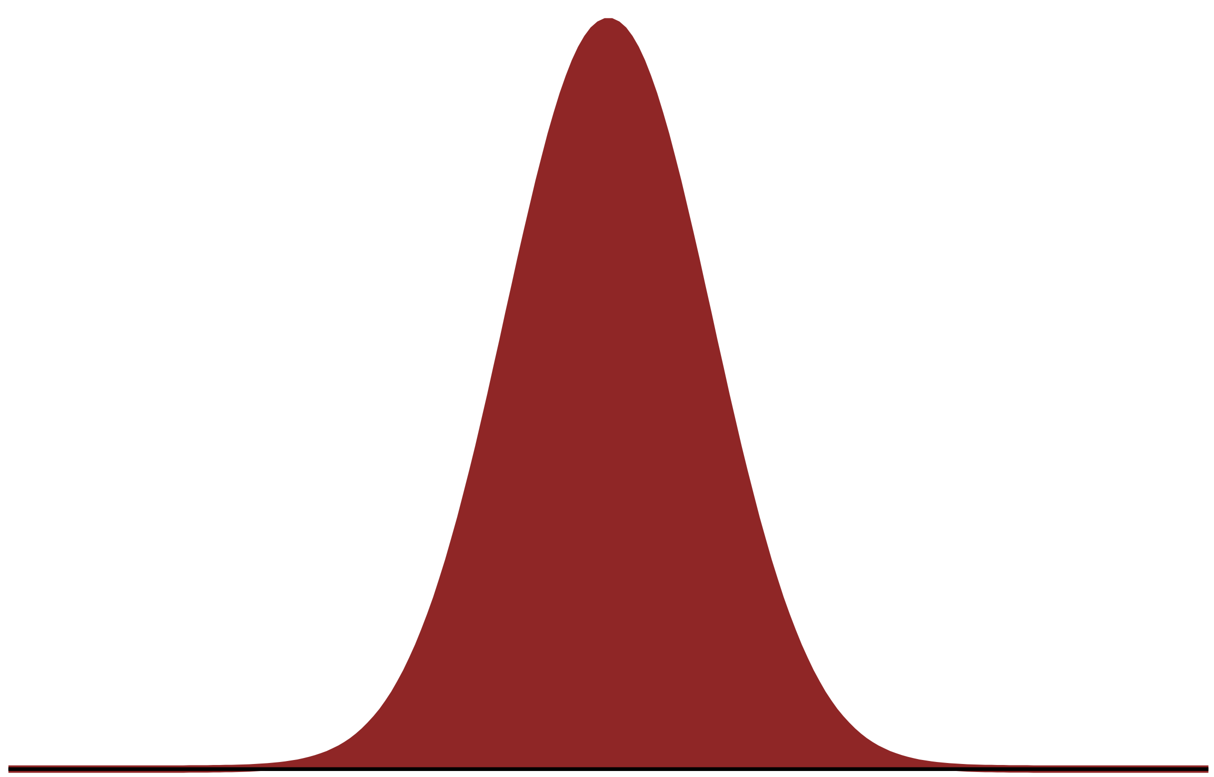
The posterior aggregates what we knew and what we learned into what we know *after* the measurement.

$$\pi_S(\theta \mid \tilde{y}) = \frac{\pi_S(\tilde{y} \mid \theta)}{\pi_S(\tilde{y})} \pi_S(\theta)$$

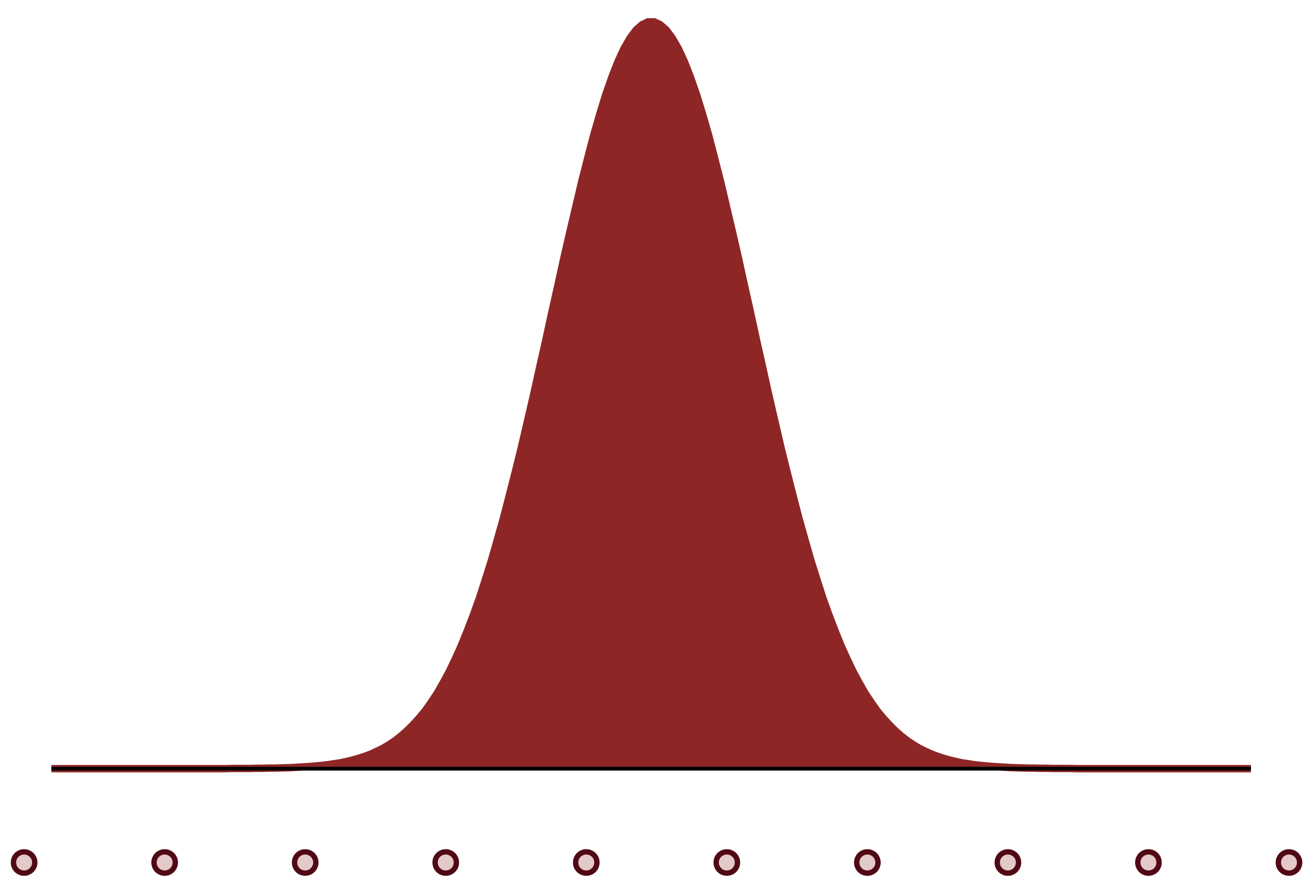
Bayesian computation concerns the computation of posterior *expectation values* once we've specified a model.

$$\mathbb{E}_{\pi}[f] = \int \mathrm{d}\theta \, \pi_{S}(\theta \mid \tilde{y}) \, f(\theta)$$

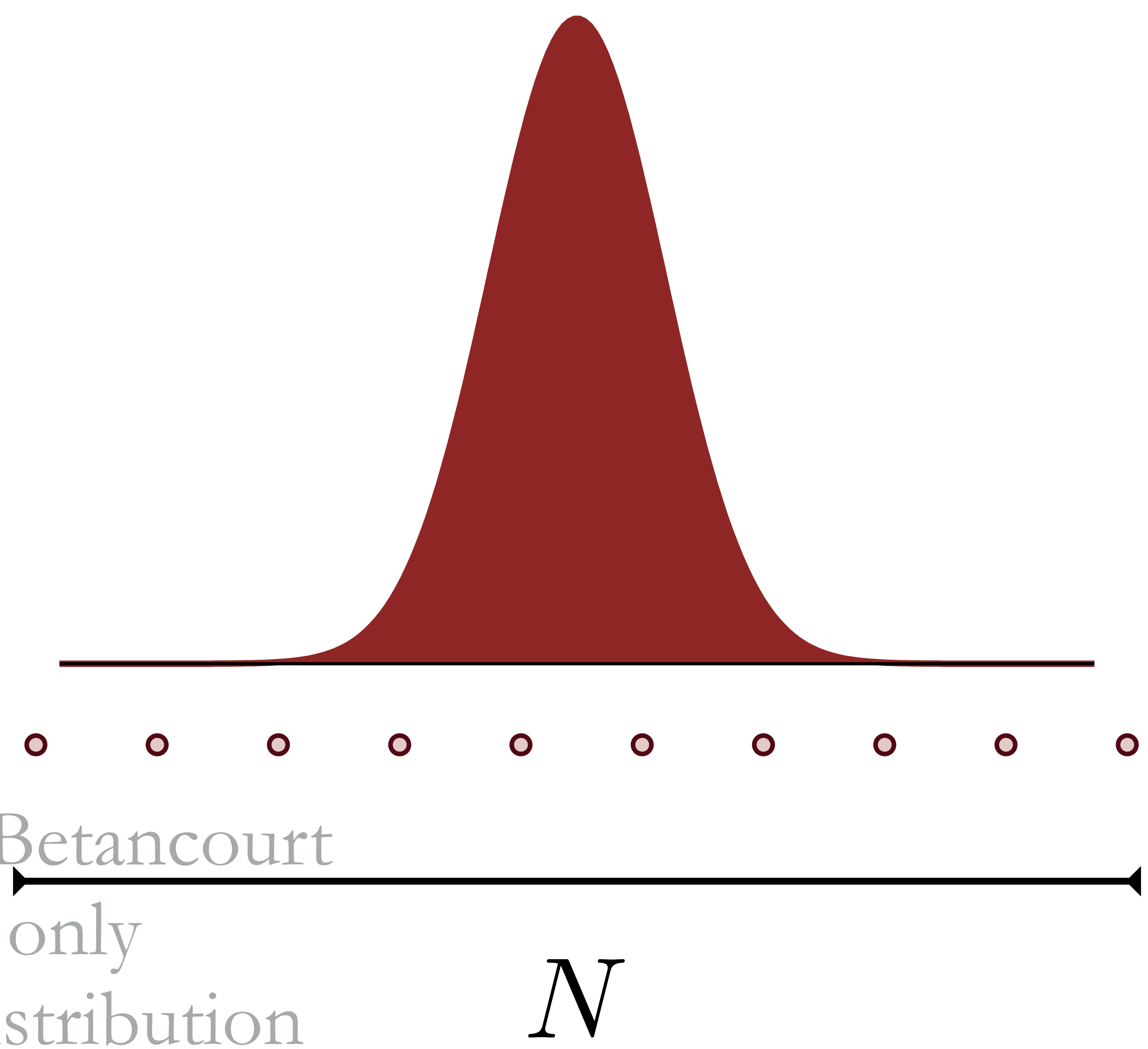
Numerical integration often takes a *quadrature* approach that mimics Riemann integration.



Numerical integration often takes a *quadrature* approach that mimics Riemann integration.

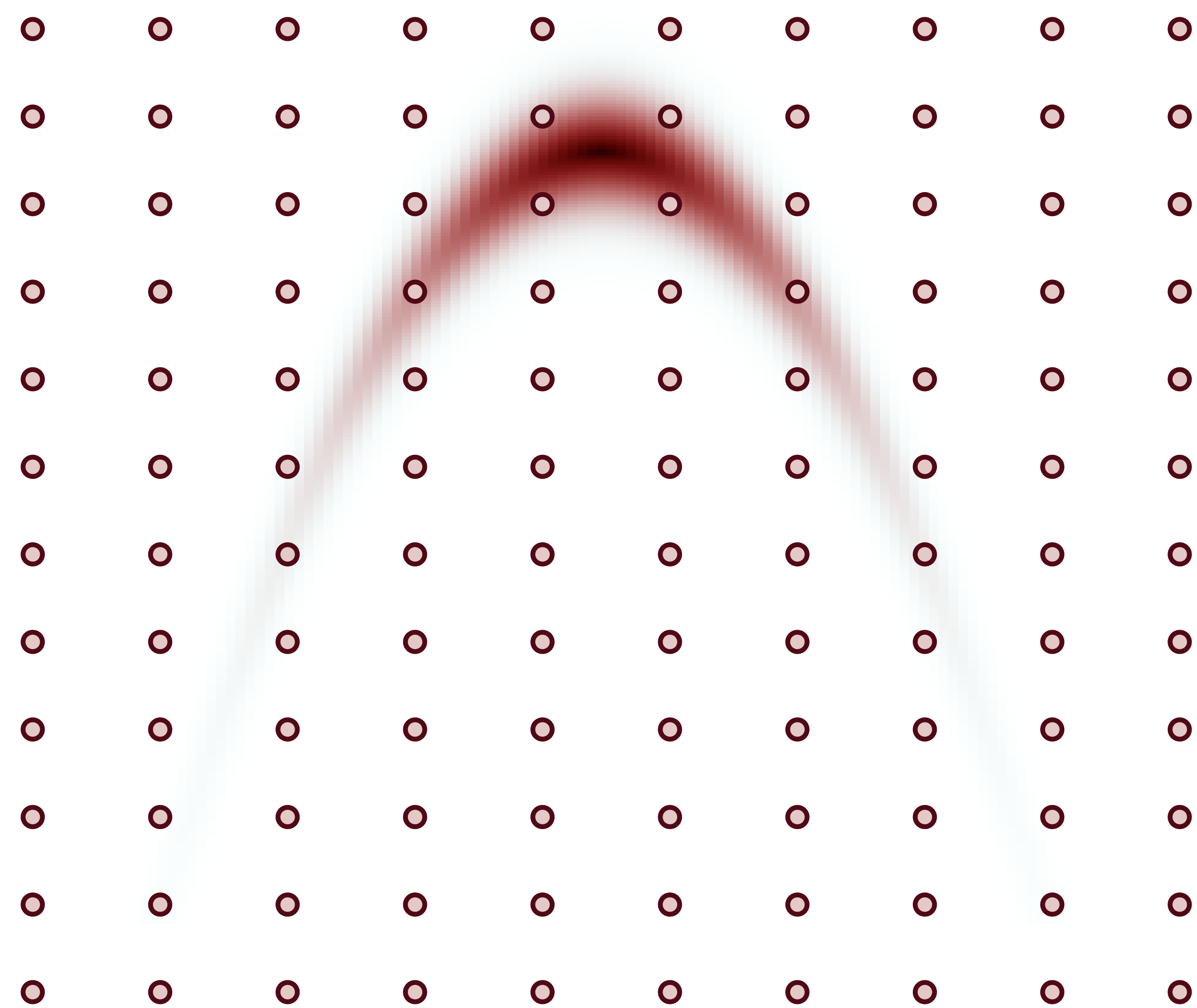


Numerical integration often takes a *quadrature* approach that mimics Riemann integration.

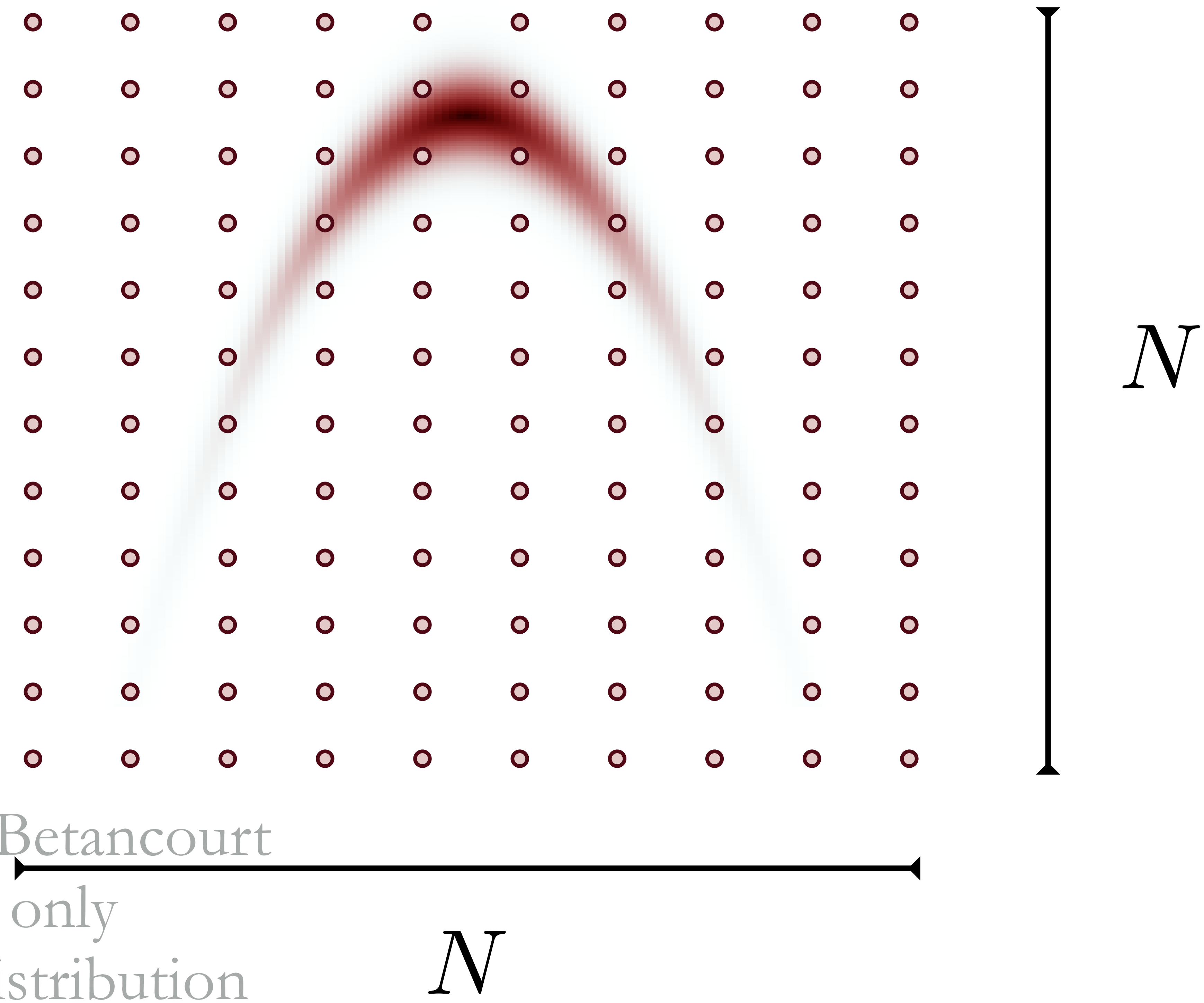


Unfortunately, the cost of numerical quadrature scales *exponentially* with the dimension of the parameter space.

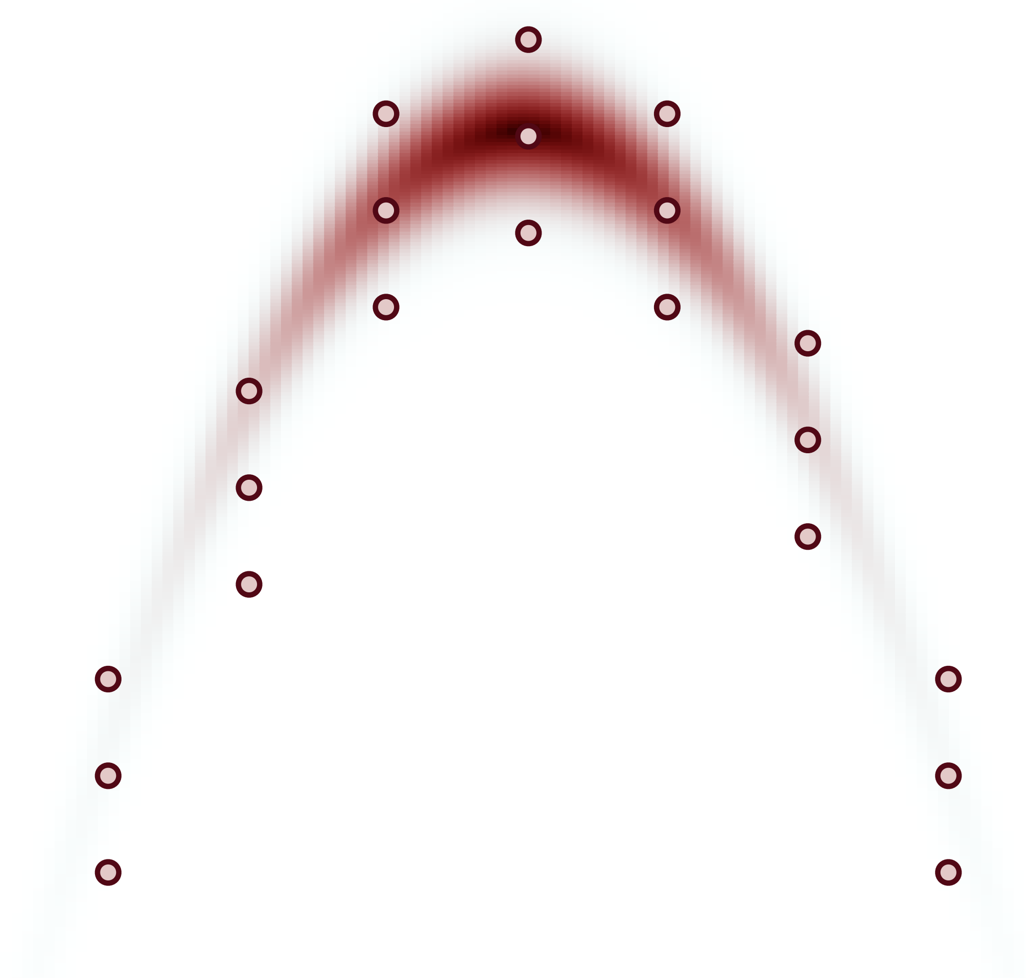
Unfortunately, the cost of numerical quadrature scales *exponentially* with the dimension of the parameter space.



Unfortunately, the cost of numerical quadrature scales *exponentially* with the dimension of the parameter space.



For the computational to be practical we need to focus on the *relevant* neighborhoods of parameter space.



But exactly which neighborhoods end up contributing the most to expectation values of reasonable functions?

$$\mathbb{E}_{\pi}[f] = \int \mathrm{d}\theta \, \pi_{S}(\theta \mid \tilde{y}) \, f(\theta)$$

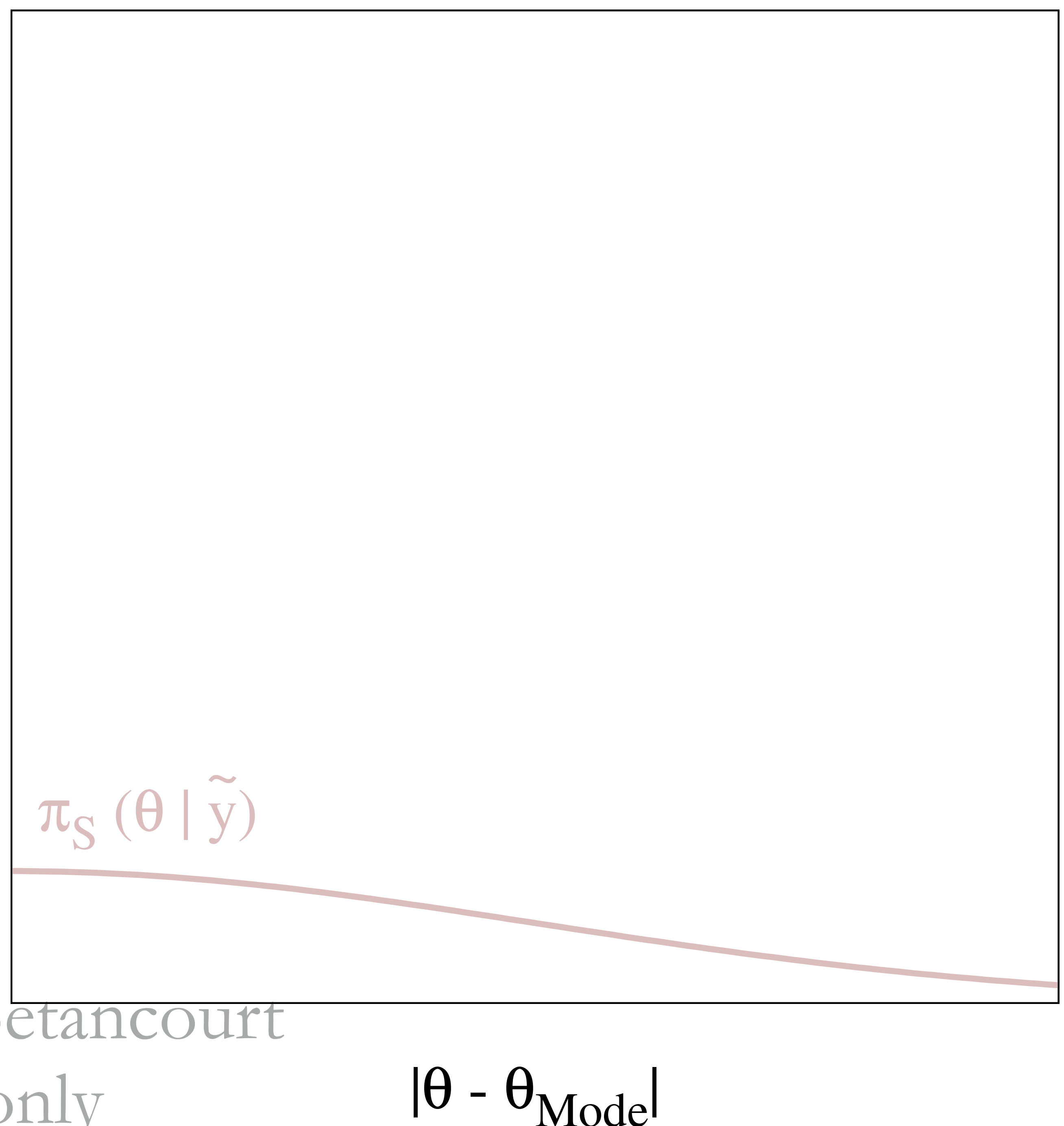
But exactly which neighborhoods end up contributing the most to expectation values of reasonable functions?

$$\mathbb{E}_{\pi}[f] = \int \mathrm{d}\theta \, \pi_{S}(\theta \mid \tilde{y}) \, f(\theta)$$

But exactly which neighborhoods end up contributing the most to expectation values of reasonable functions?

$$\mathbb{E}_{\pi}[f] = \int \mathrm{d}\theta \, \pi_{S}(\theta \mid \tilde{y}) \, f(\theta)$$

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



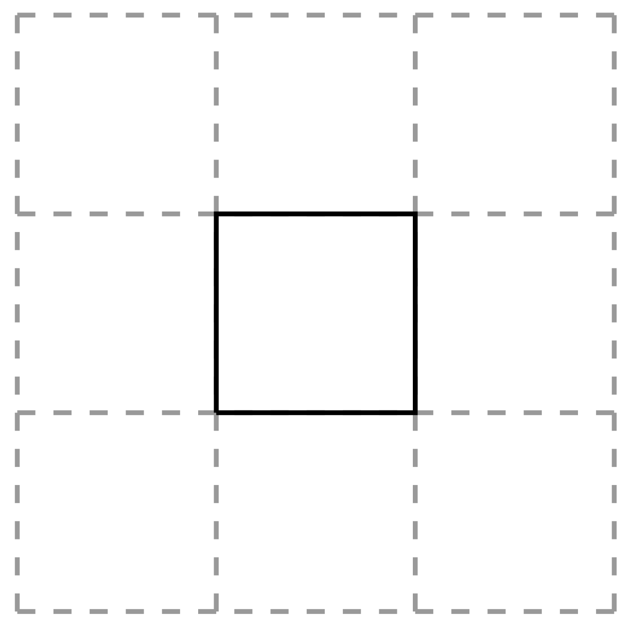
But integration doesn't just evaluate the integrand -- it aggregates it over volumes.

$$\mathbb{E}_{\pi}[f] = \int \mathrm{d}\theta \, \pi_{S}(\theta \mid \tilde{y}) \, f(\theta)$$

Volume starts to behave counterintuitively as the dimension of the parameter space increases.

1D

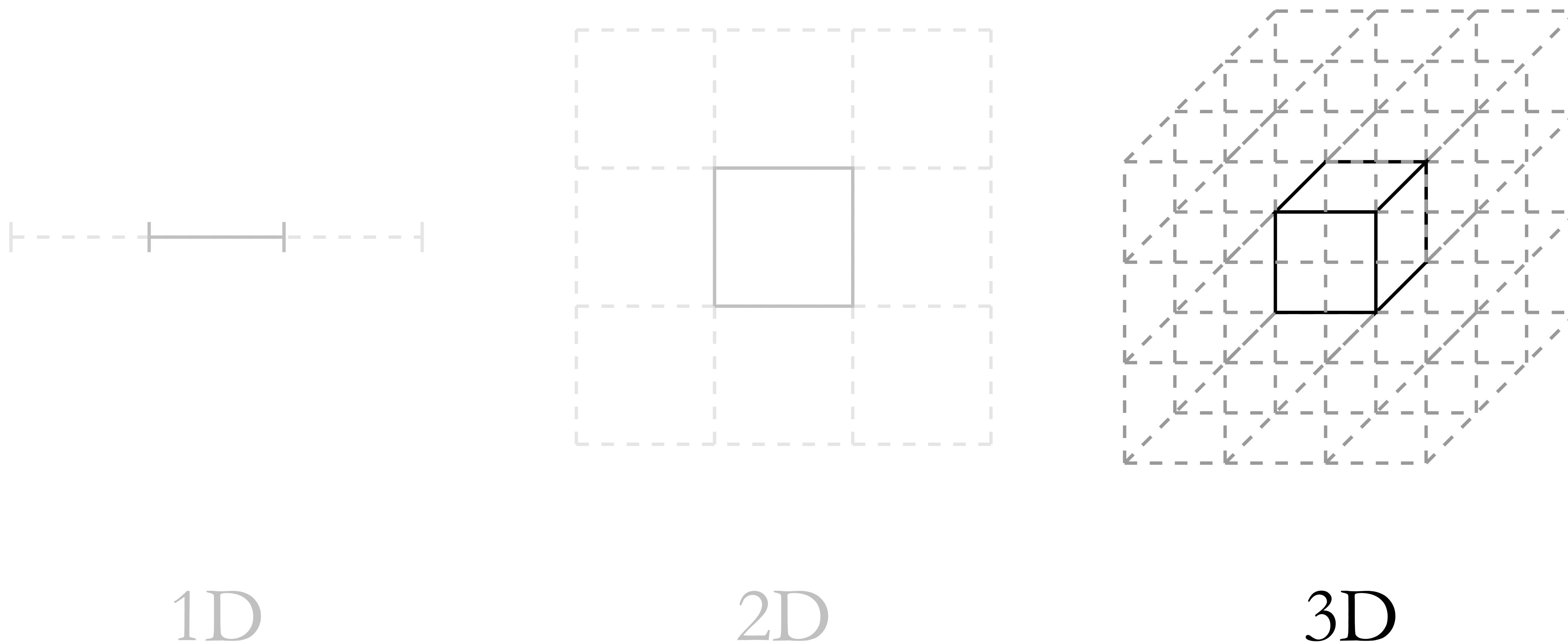
Volume starts to behave counterintuitively as the dimension of the parameter space increases.



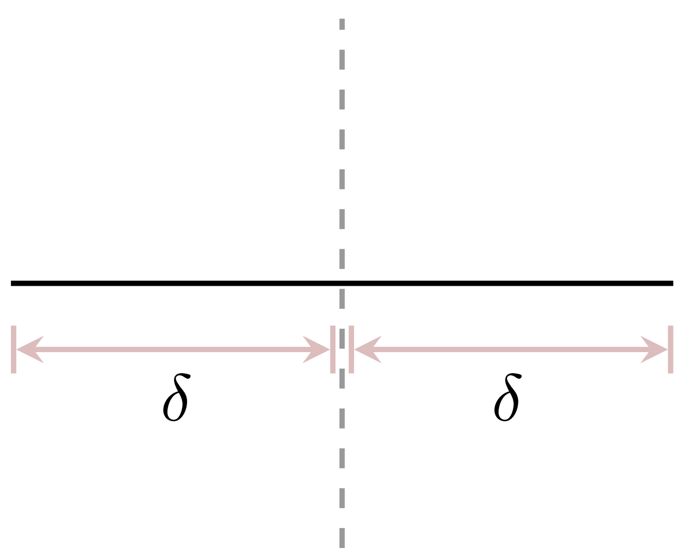
1D

2D

Volume starts to behave counterintuitively as the dimension of the parameter space increases.

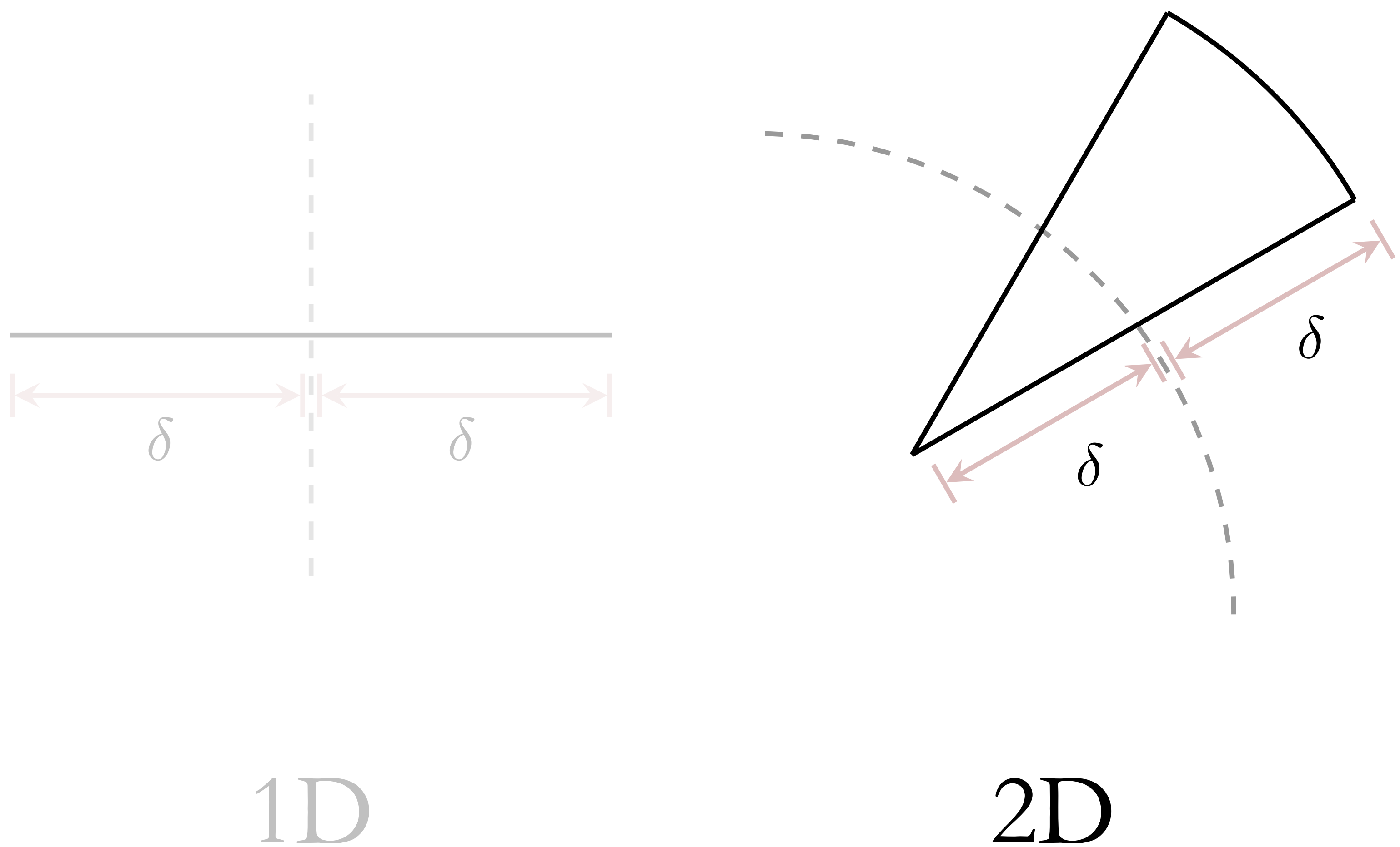


Volume starts to behave counterintuitively as the dimension of the parameter space increases.

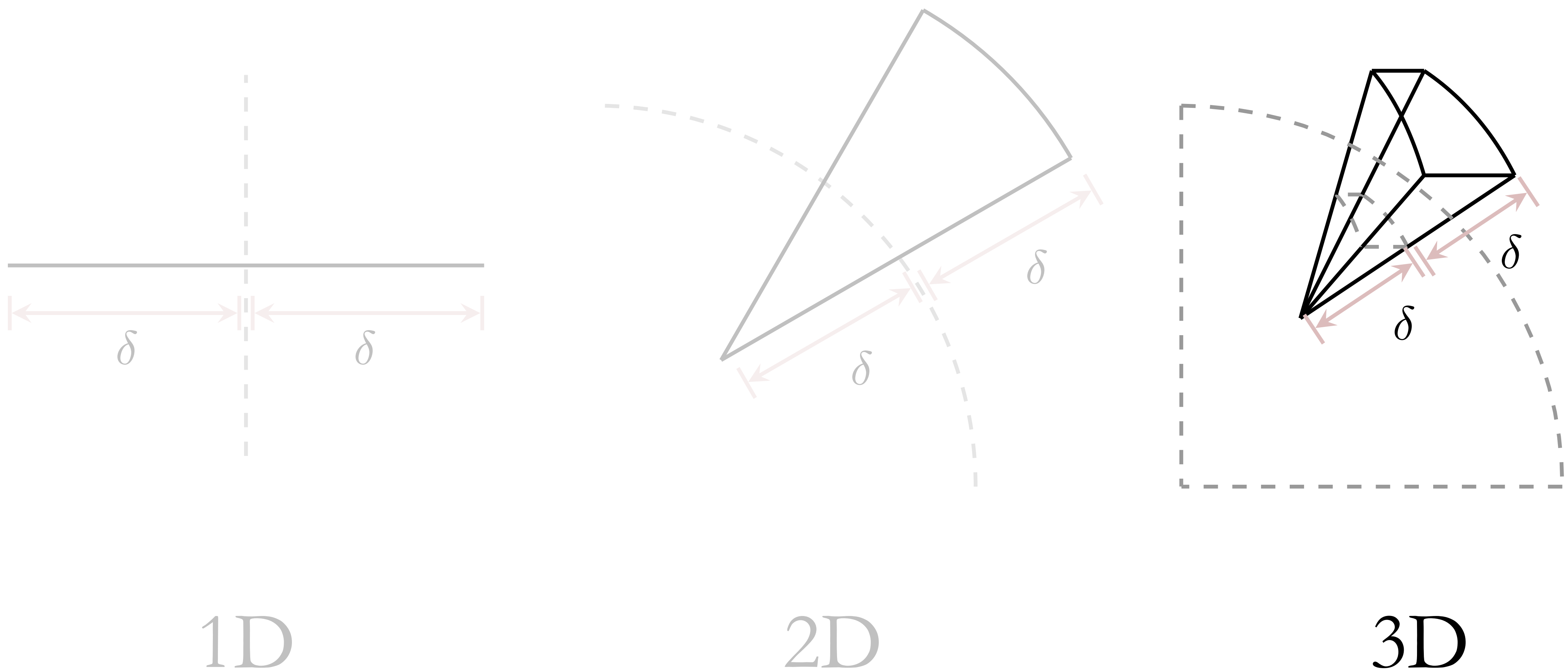


1D

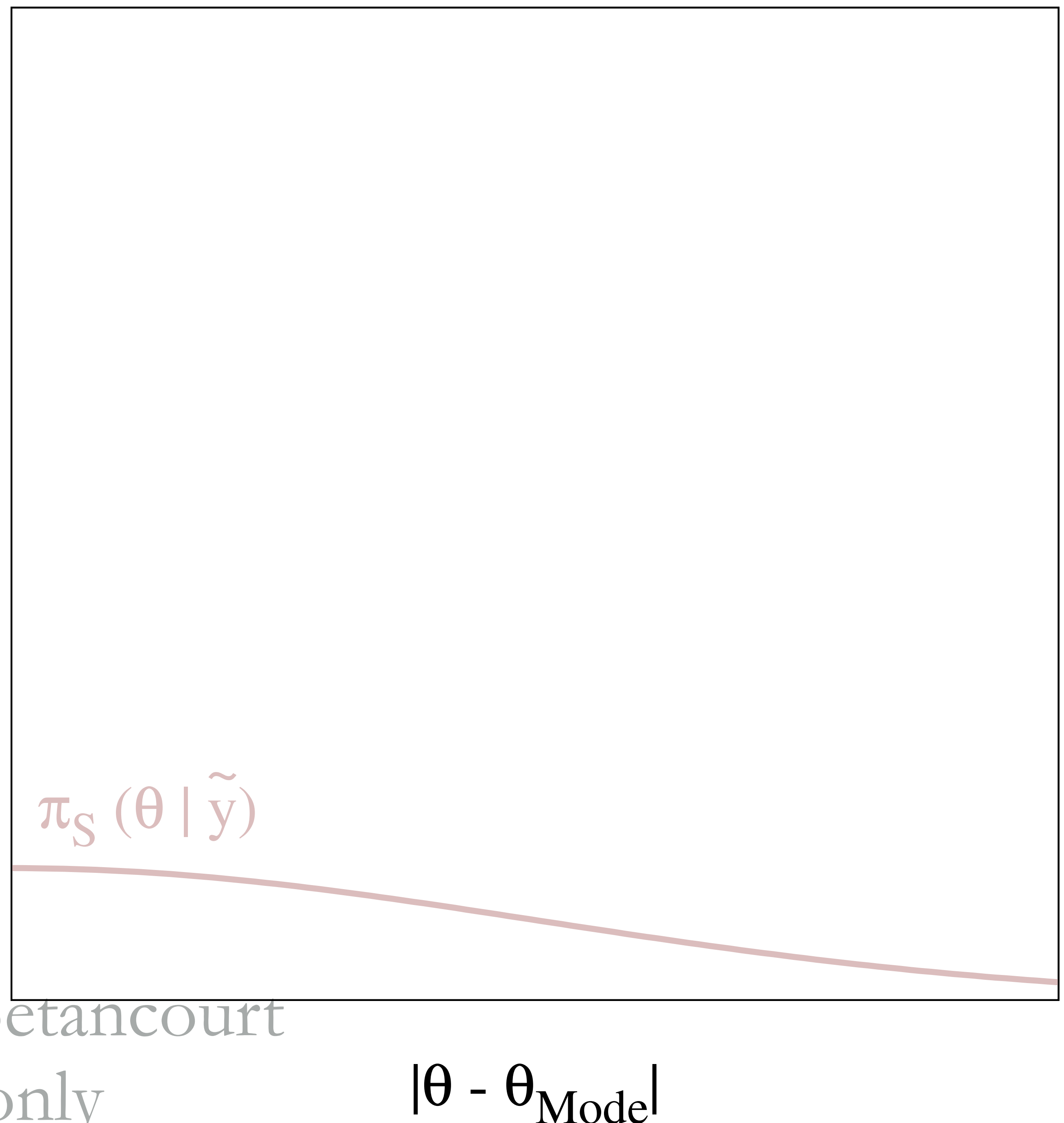
Volume starts to behave counterintuitively as the dimension of the parameter space increases.



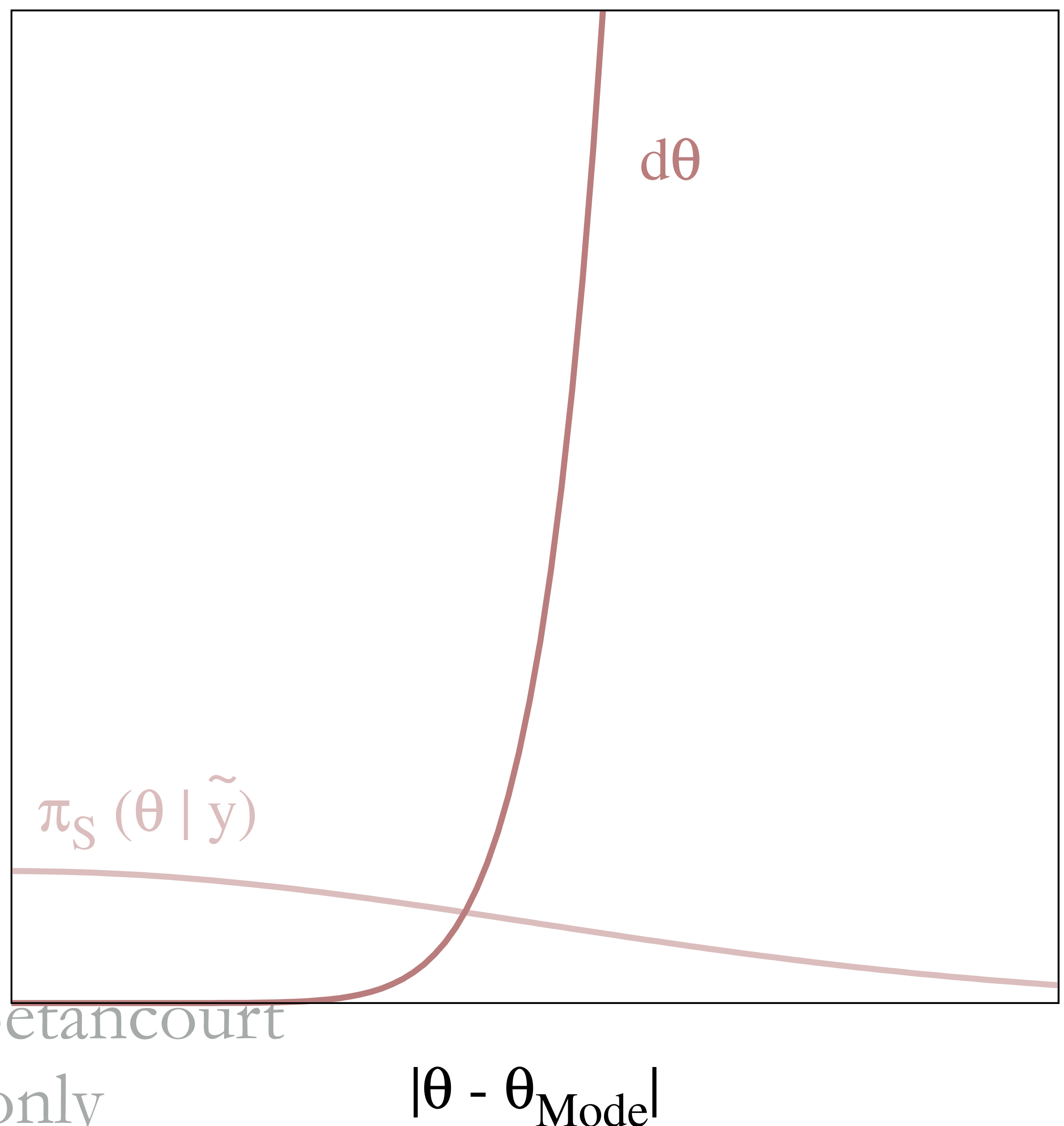
Volume starts to behave counterintuitively as the dimension of the parameter space increases.



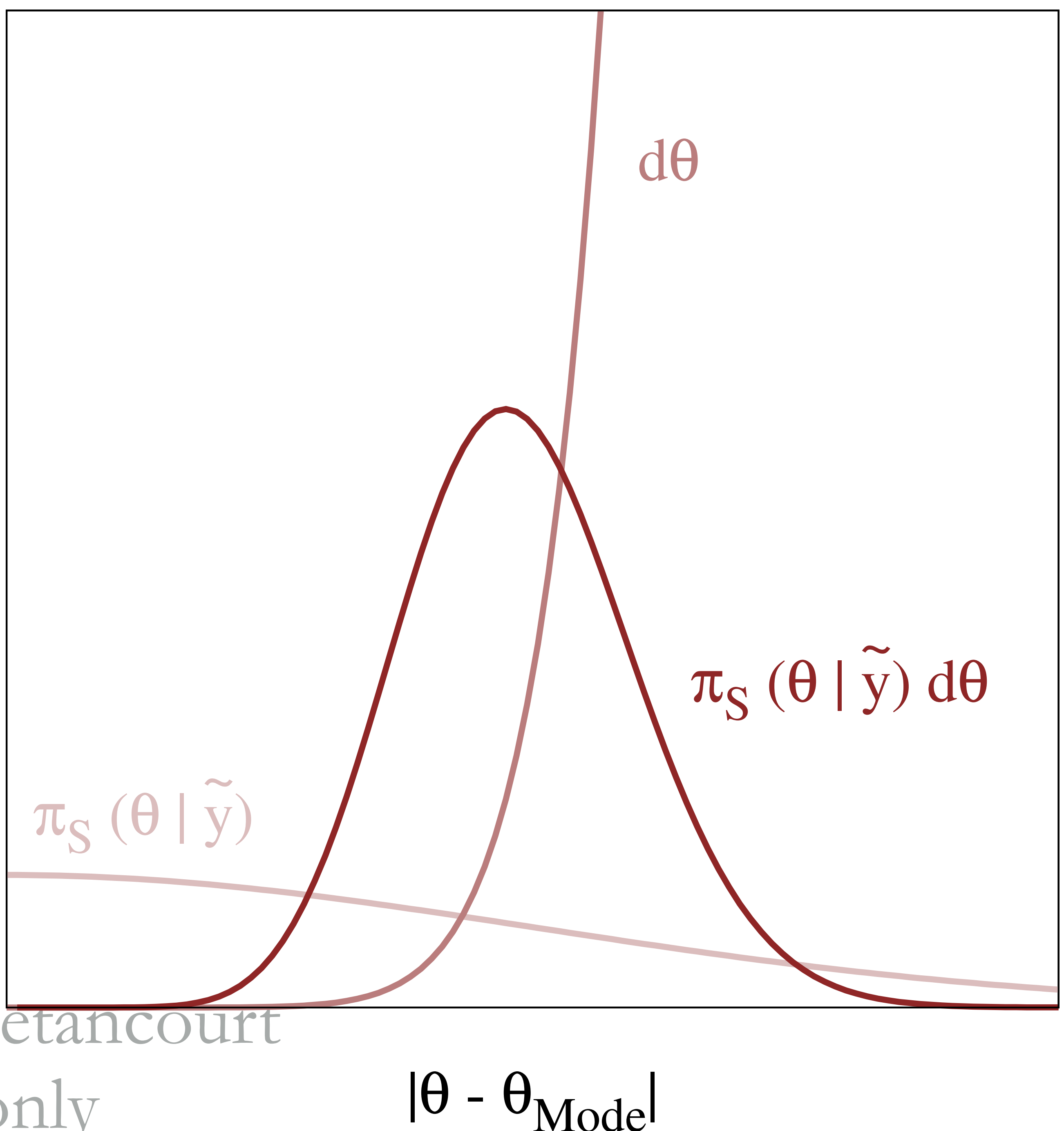
The dominant contributions to these integrals are dictated not by probability *density* but rather by probability *mass*.



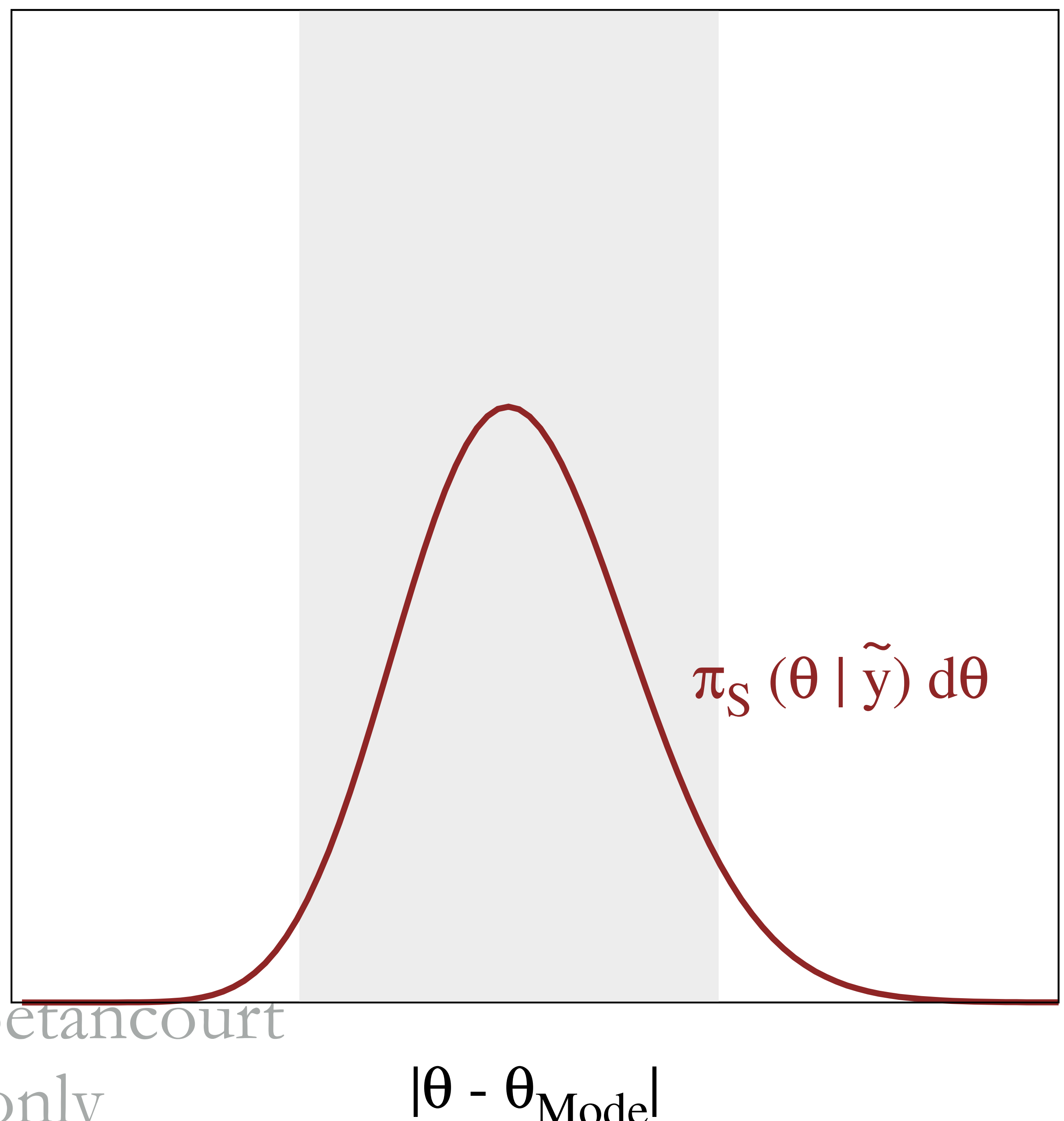
The dominant contributions to these integrals are dictated not by probability *density* but rather by probability *mass*.



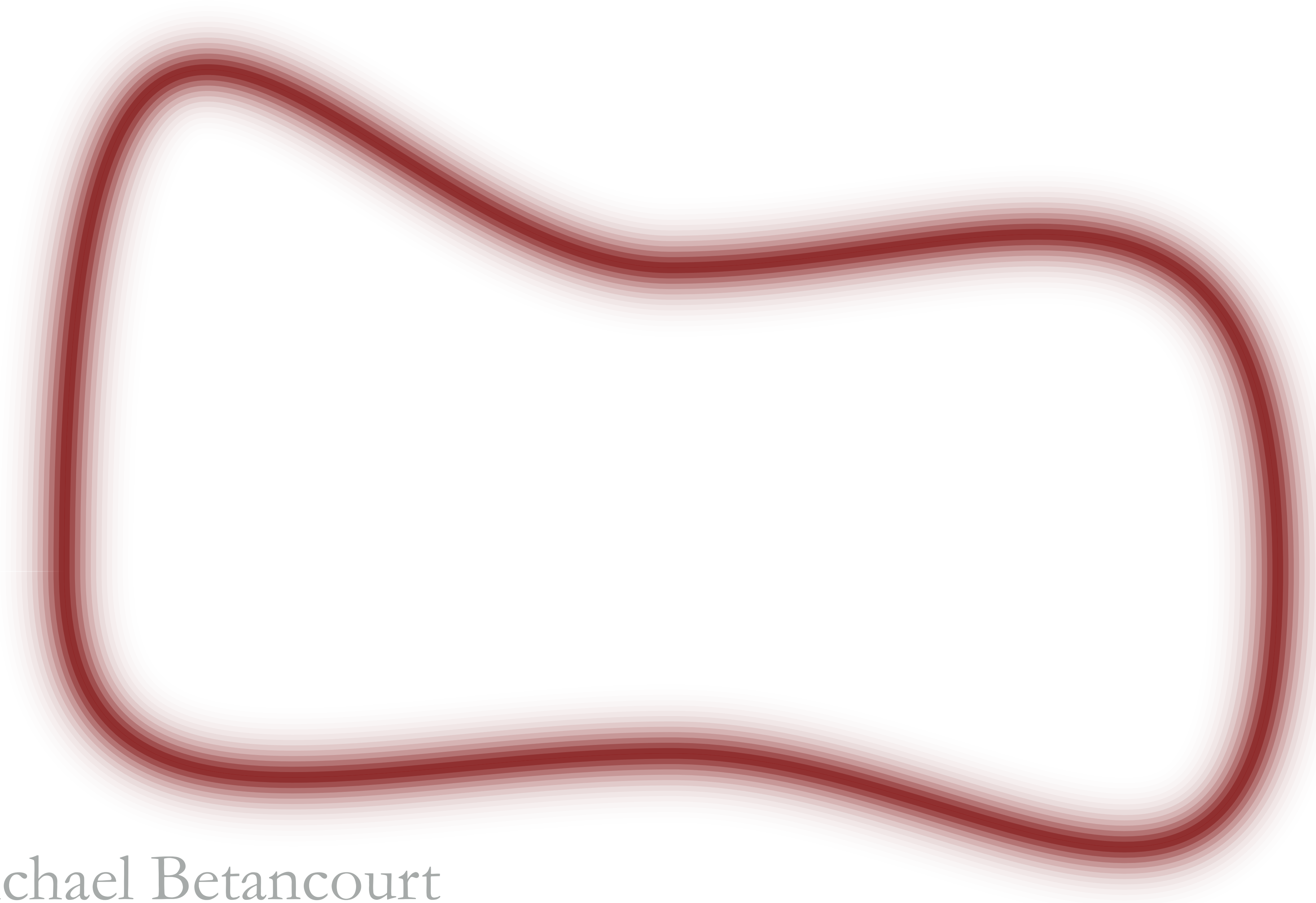
The dominant contributions to these integrals are dictated not by probability *density* but rather by probability *mass*.



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



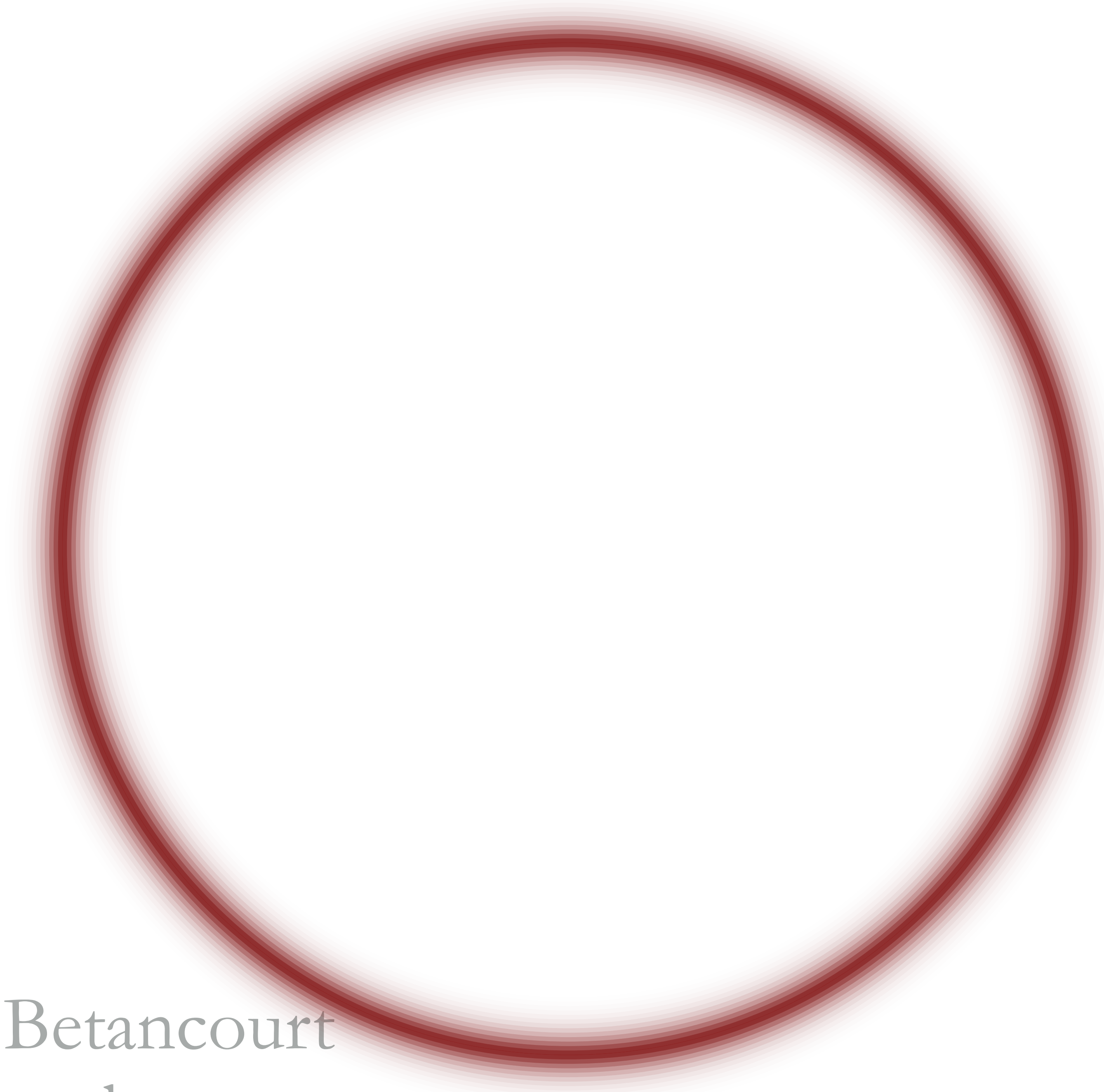
This *concentration of measure* into a narrow typical set frustrates the accurate estimation of integrals.



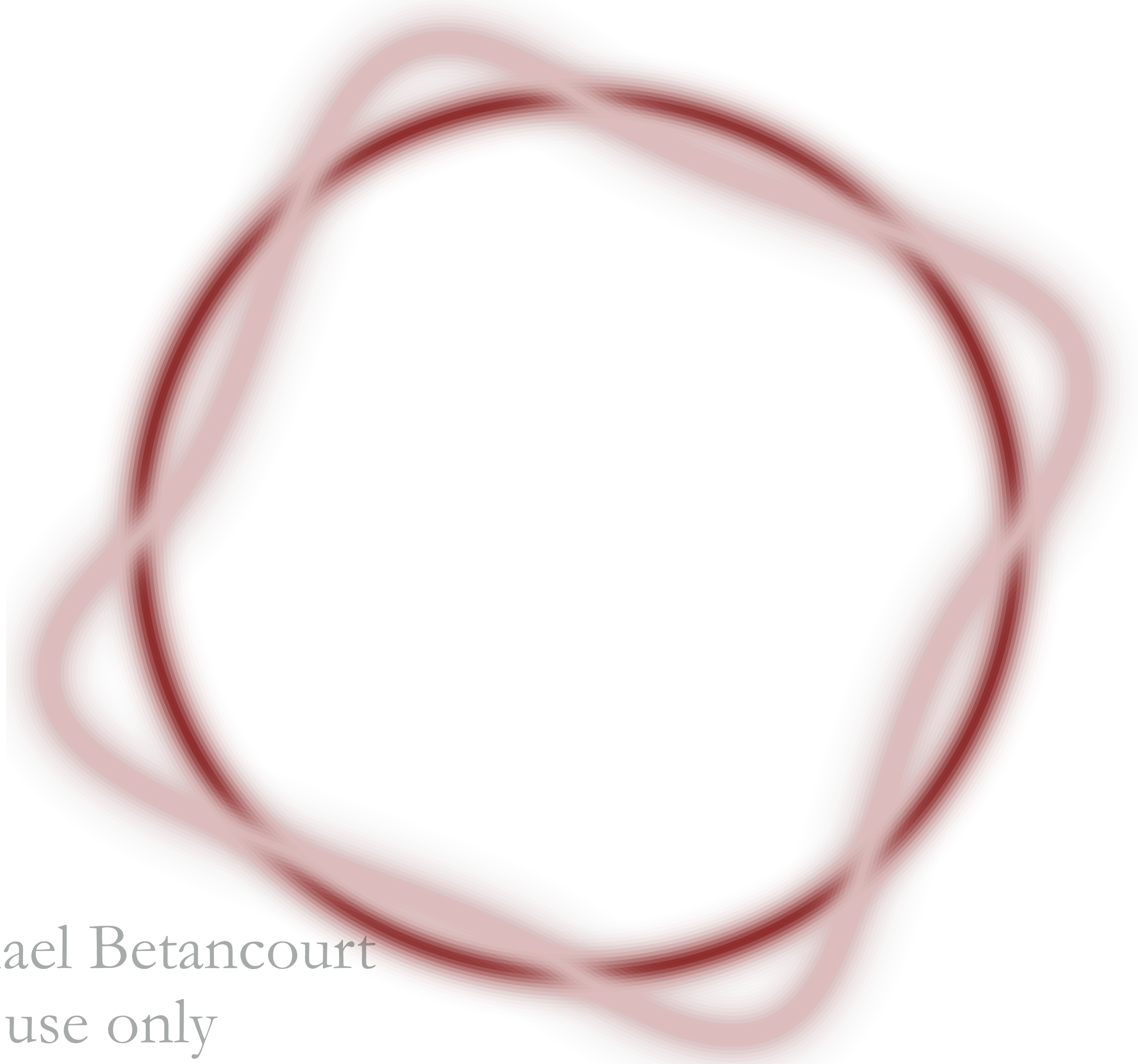
Any method that accurately estimates expectation values is really a method of quantifying the typical set.

Deterministic	*Stochastic*
Modal Estimators	Importance Sampling
Laplace Estimators	Monte Carlo
Variational Estimators	Markov Chain Monte Carlo
...	...

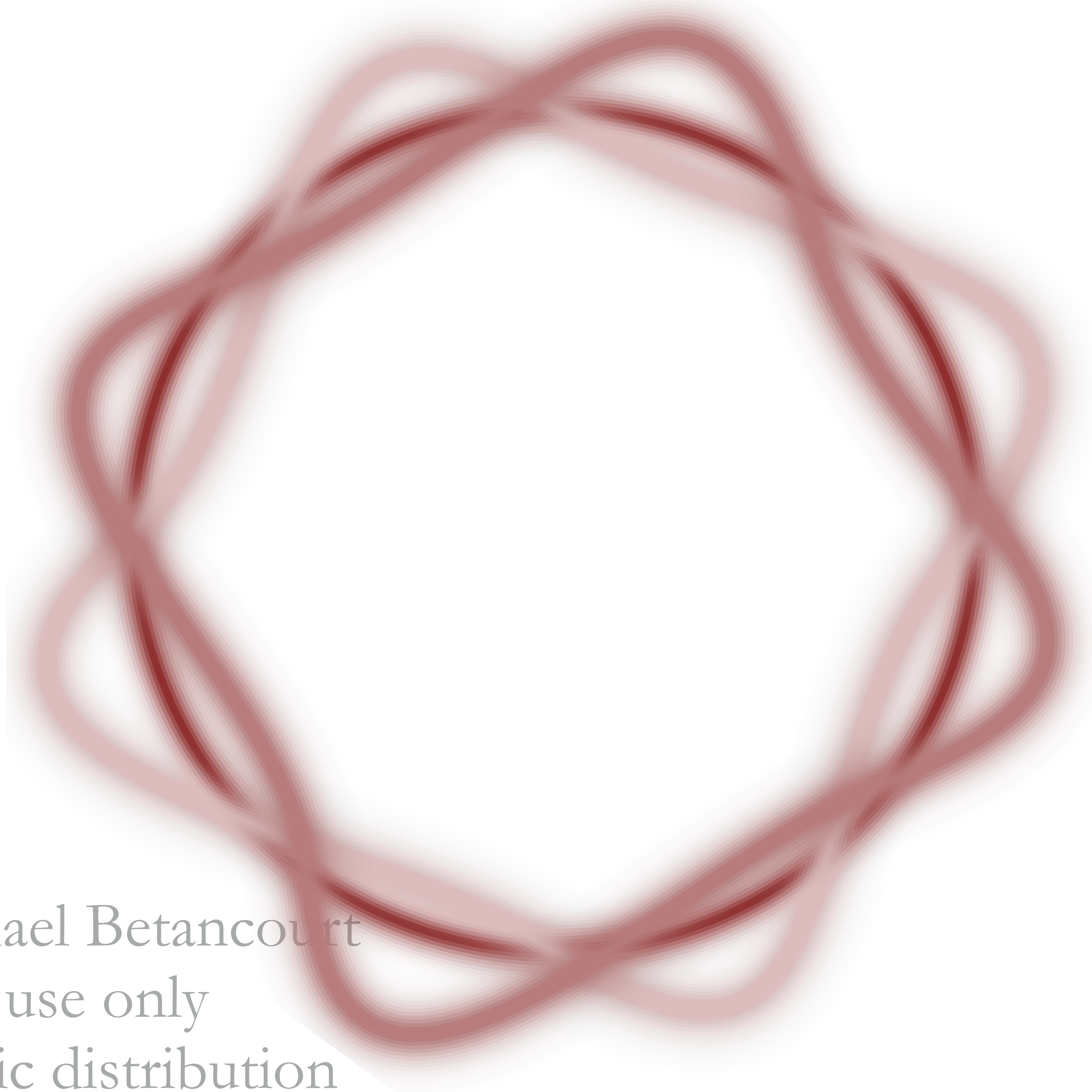
This perspective facilitates understanding the often unfortunate consequences of various approaches.



This perspective facilitates understanding the often unfortunate consequences of various approaches.



This perspective facilitates understanding the often unfortunate consequences of various approaches.



We can also use it to analyze particular algorithms and build intuition about their robustness, or lack thereof.

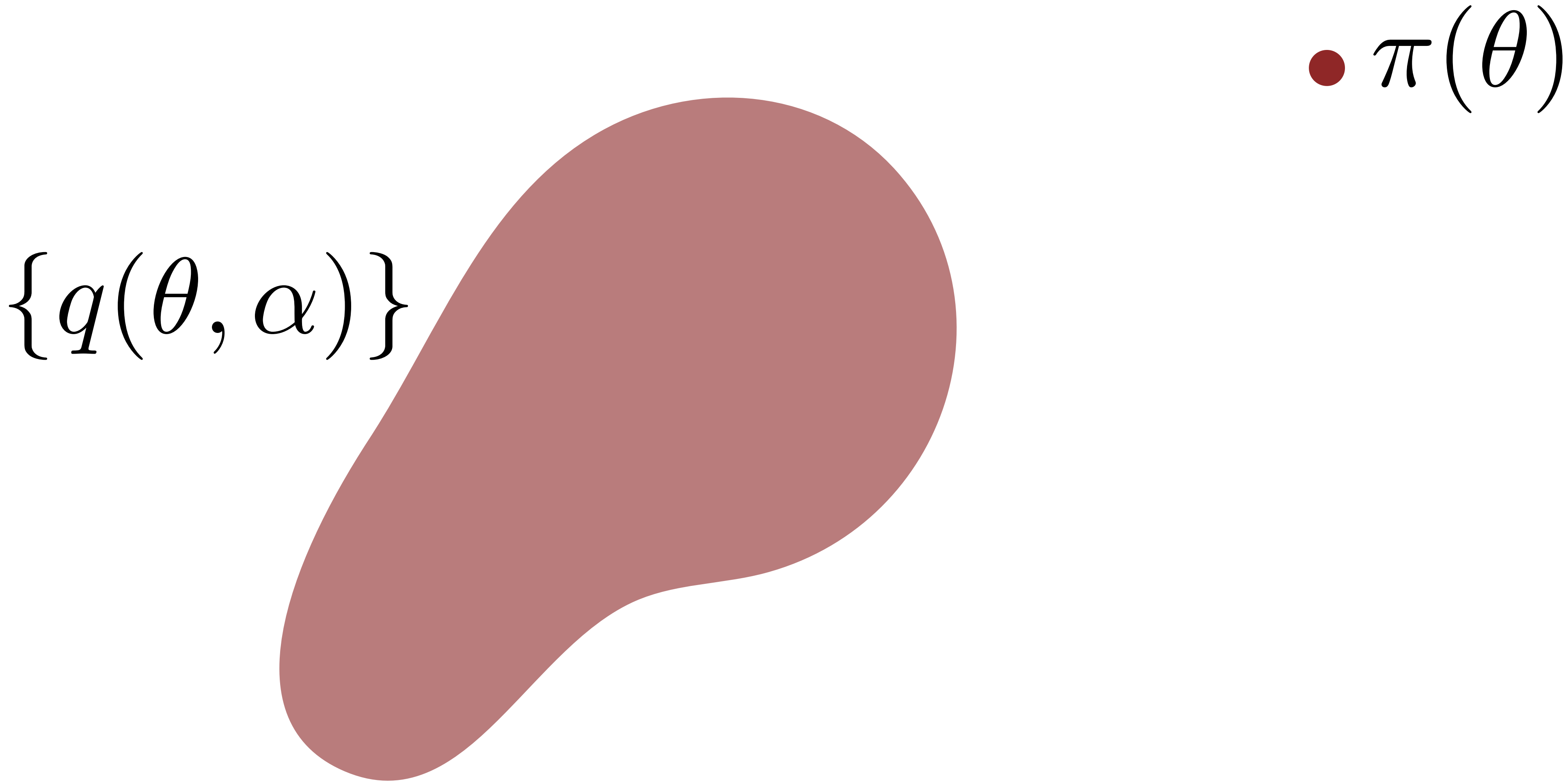
Deterministic
Modal Estimators
Laplace Estimators
Variational Estimators
...

Stochastic
Rejection Sampling
Importance Sampling
Markov Chain Monte Carlo
...

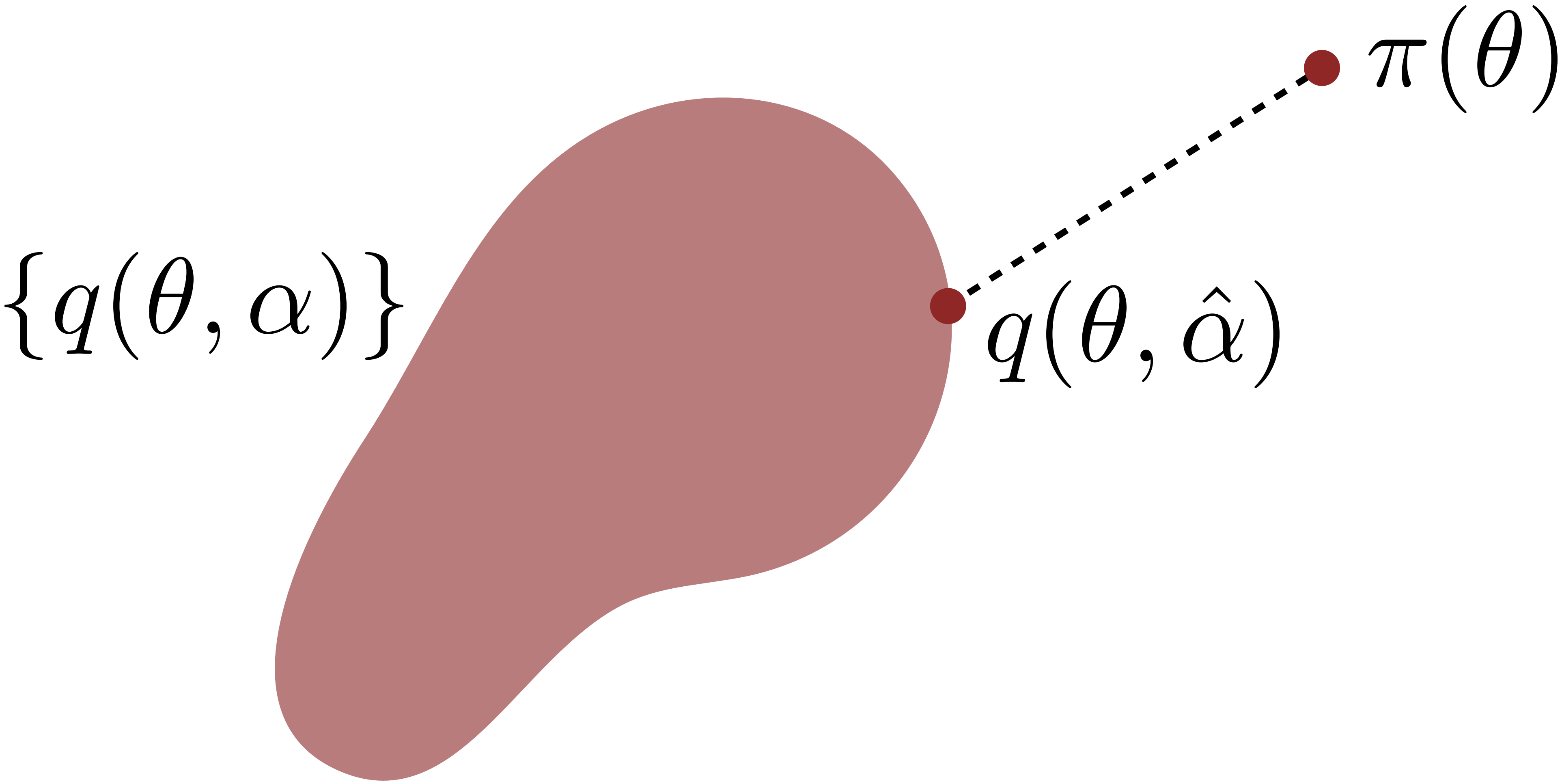
Variational methods try to optimize over a given, often convenient, family of approximating distributions.

$\bullet\ \pi(\theta)$

Variational methods try to optimize over a given, often convenient, family of approximating distributions.



Variational methods try to optimize over a given, often convenient, family of approximating distributions.

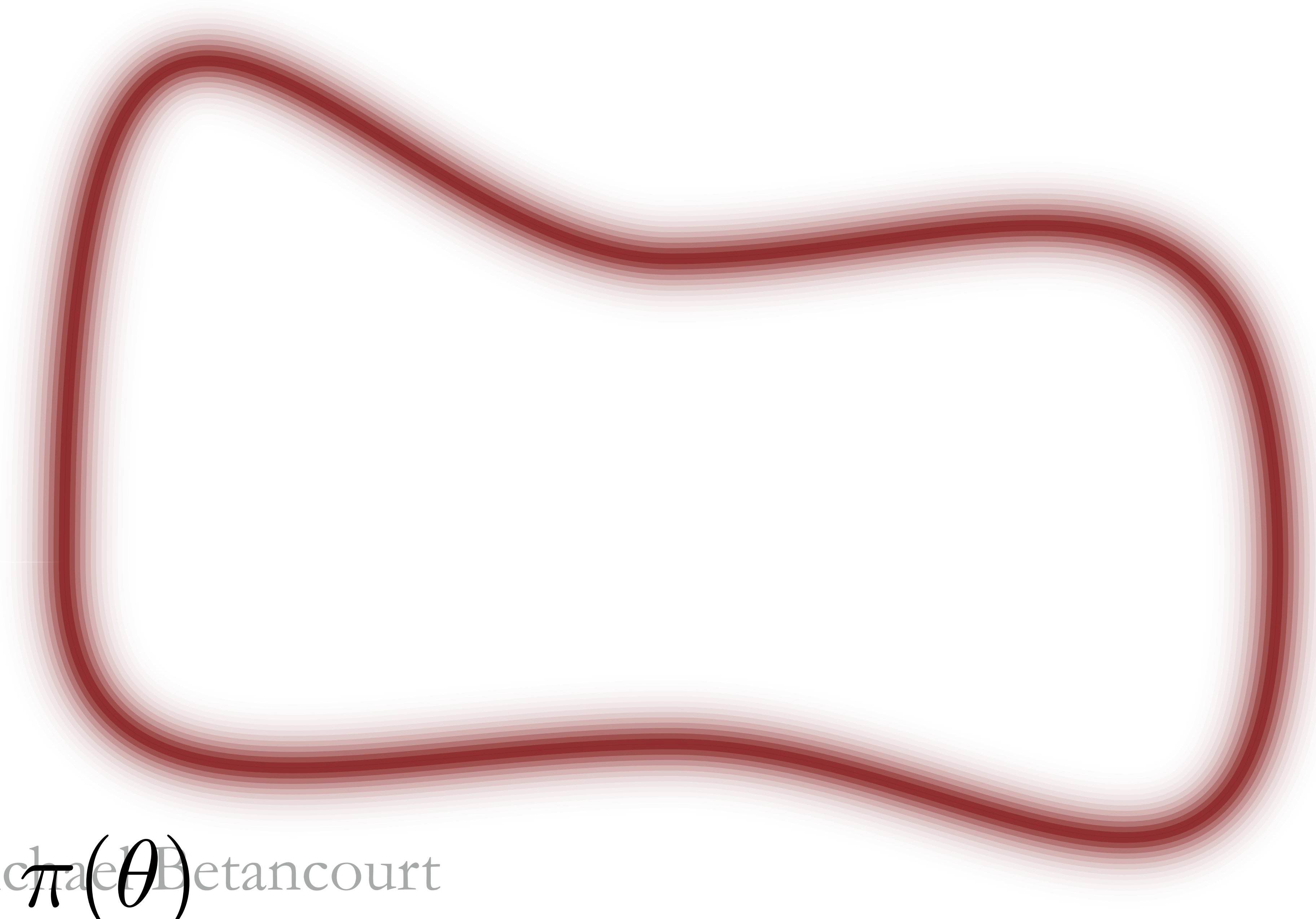


The local variational solution is then used to approximate the target expectation values.

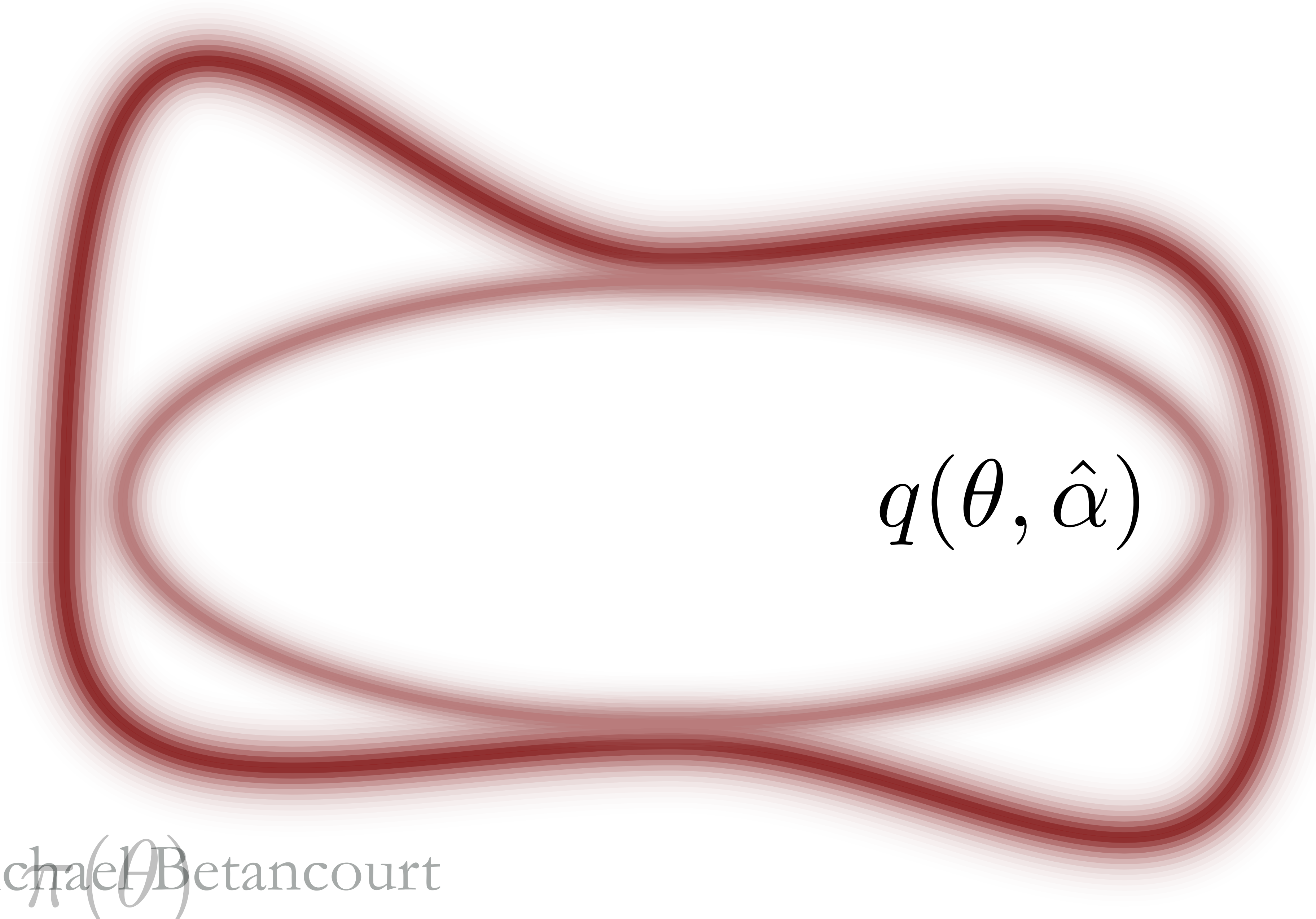
$$\pi(\theta) \approx q(\theta, \hat{\alpha})$$

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

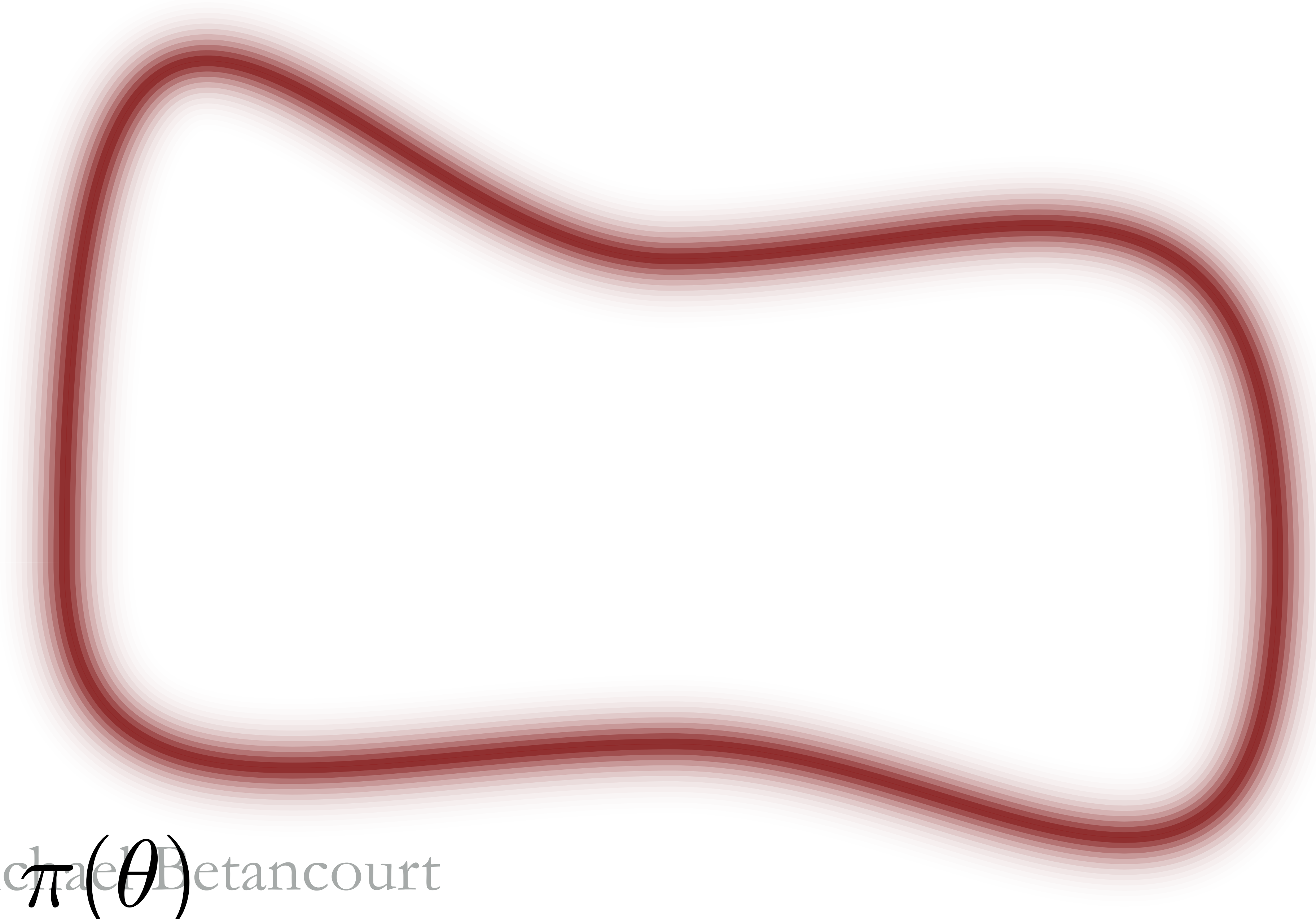
The variational objective function in "VI" methods favors solutions whose typical sets fall *inside* the target typical set.



The variational objective function in "VI" methods favors solutions whose typical sets fall *inside* the target typical set.



Other variational objective functions can favor solutions whose typical sets fall *outside* the target typical set.



Other variational objective functions can favor solutions whose typical sets fall *outside* the target typical set.