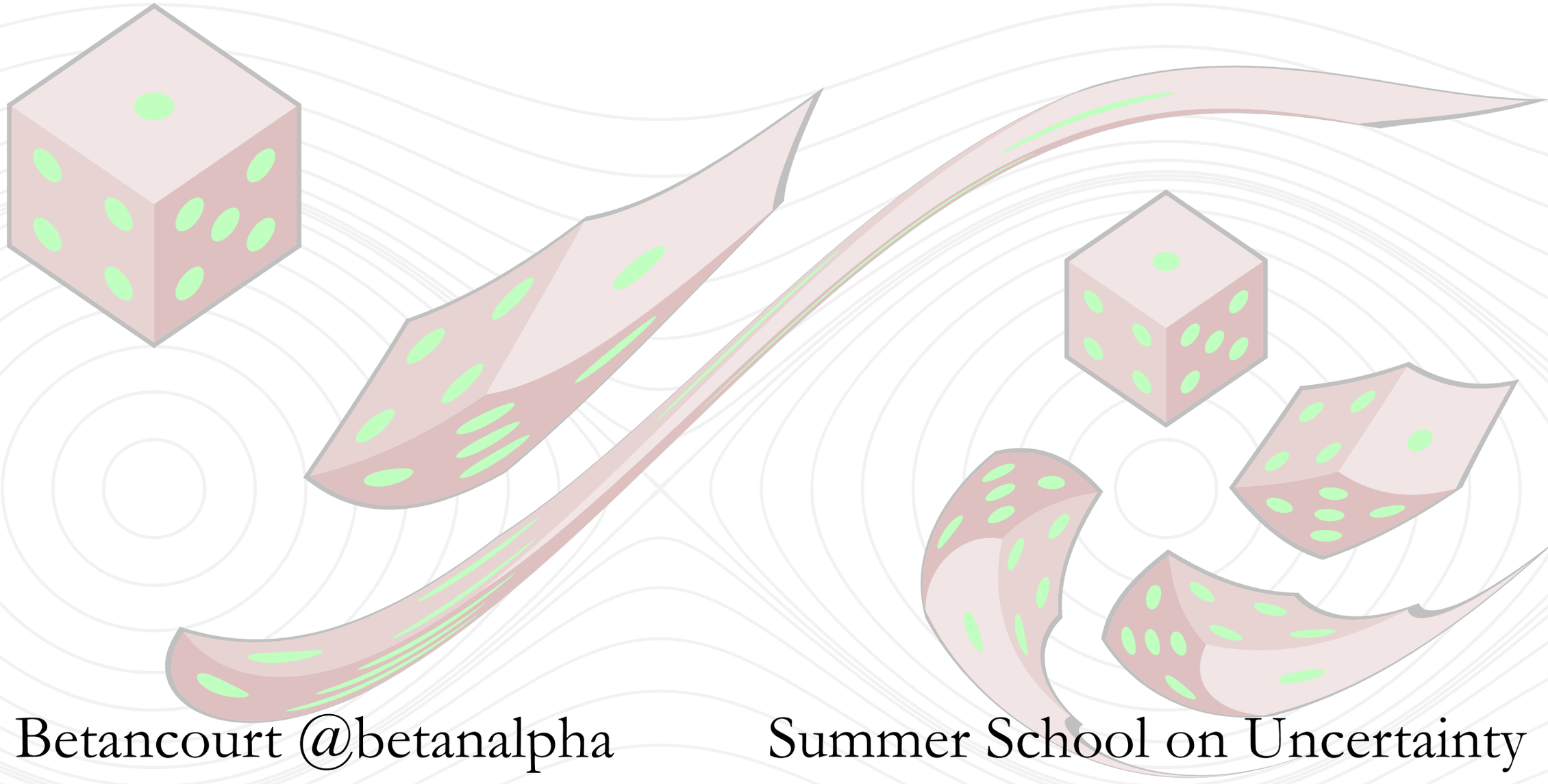
Scalable Bayesian Inference with Hamiltonian Monte Carlo



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

Summer School on Uncertainty
Quantification for Applied Problems,
BCAM, Bilbao, Spain
July 4, 2016

Foundations of Inference and Computation

Foundations of Inference and Computation

Markov Chain Monte Carlo and Hamiltonian Monte Carlo

Foundations of Inference and Computation

Markov Chain Monte Carlo and Hamiltonian Monte Carlo

Regression Modeling

Foundations of Inference and Computation

Markov Chain Monte Carlo and
Hamiltonian Monte Carlo

Regression Modeling

Hierarchical Modeling

Foundations of Bayesian Inference

Attempts to learn from measurements is complicated by the natural variability of measurements.

$\mathcal{D}$

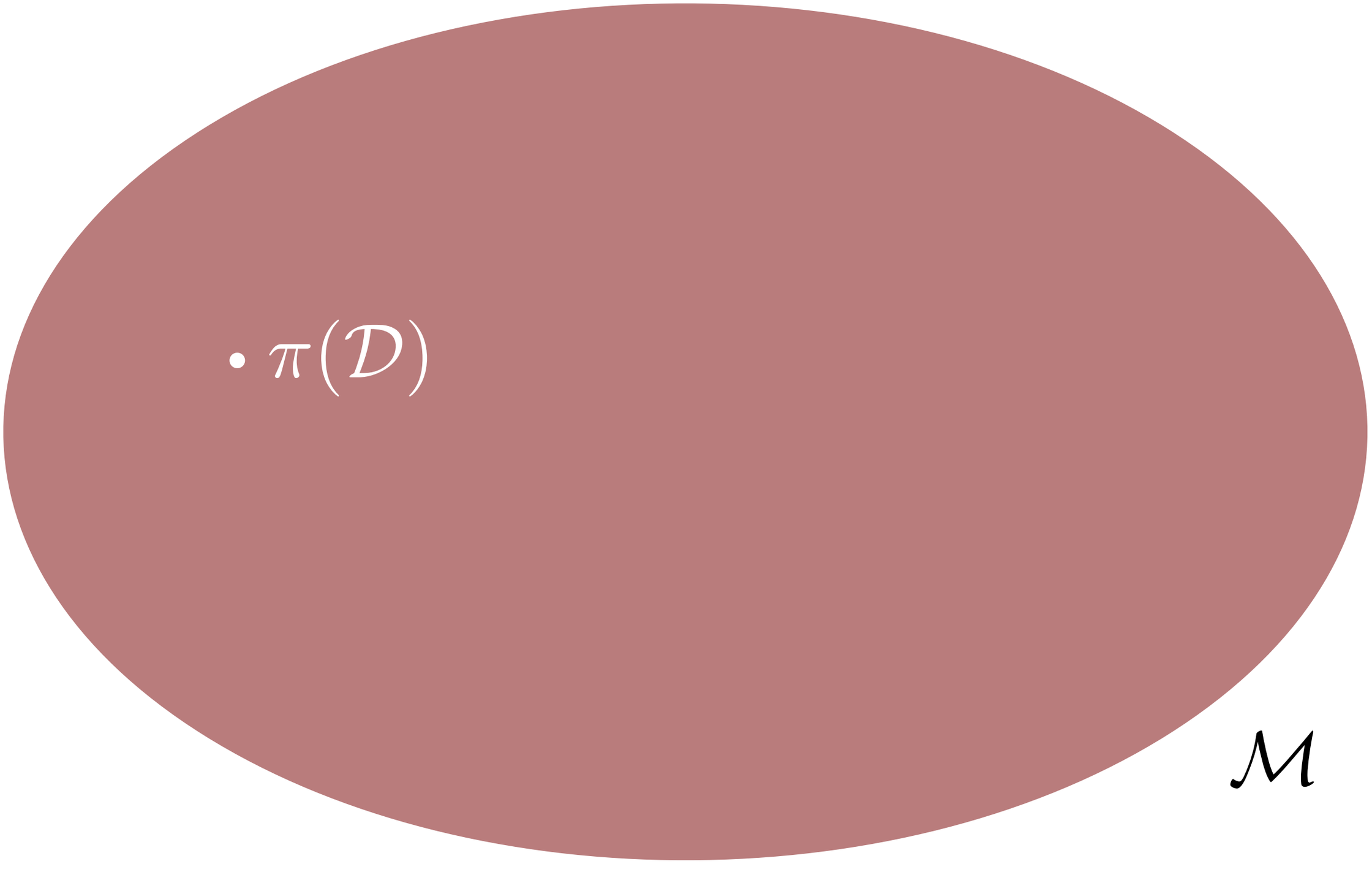
Attempts to learn from measurements is complicated by the natural variability of measurements.

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

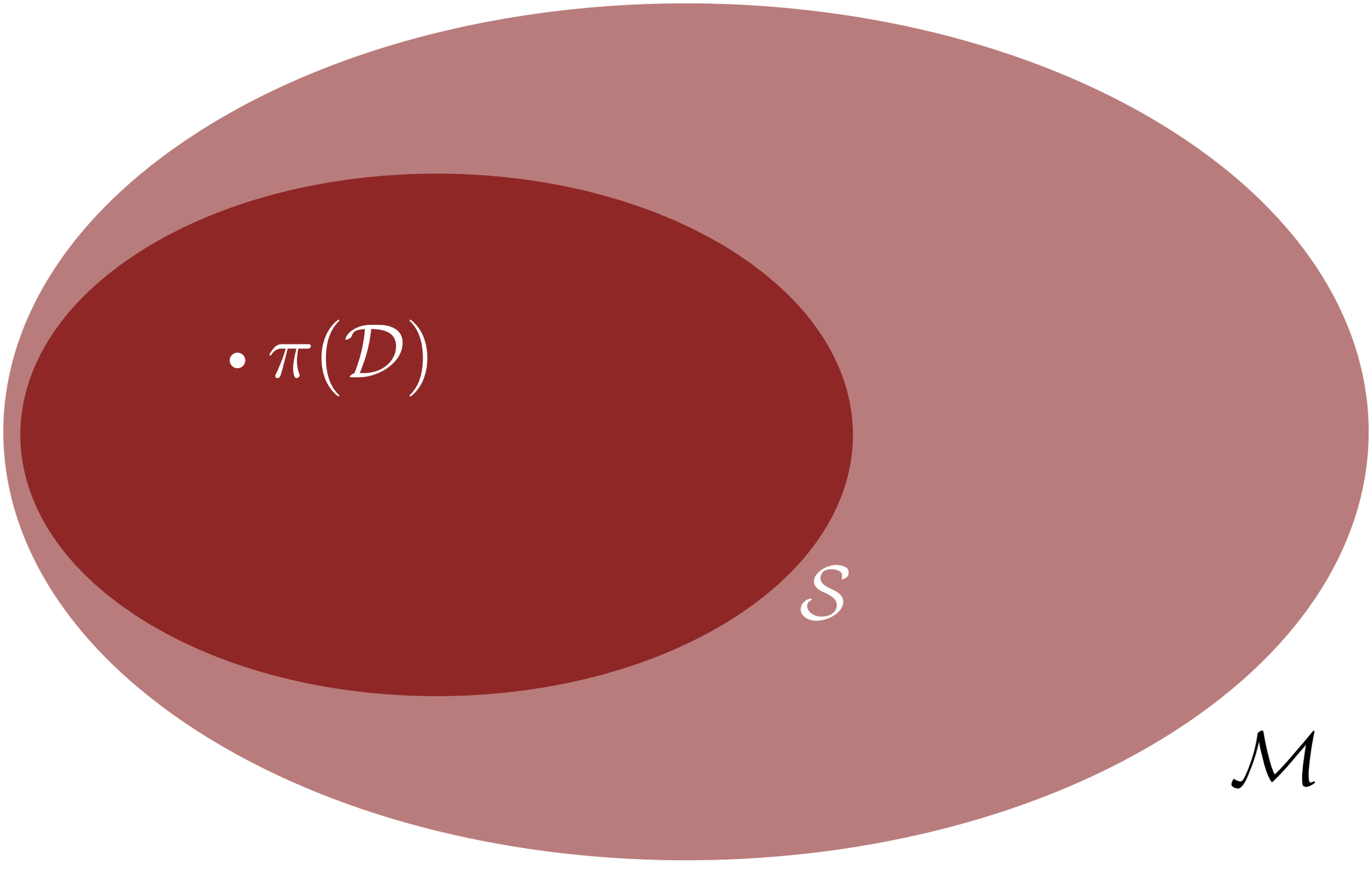
Ultimately all inference assumes the existence of, and attempts to model, some latent data generating process.

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

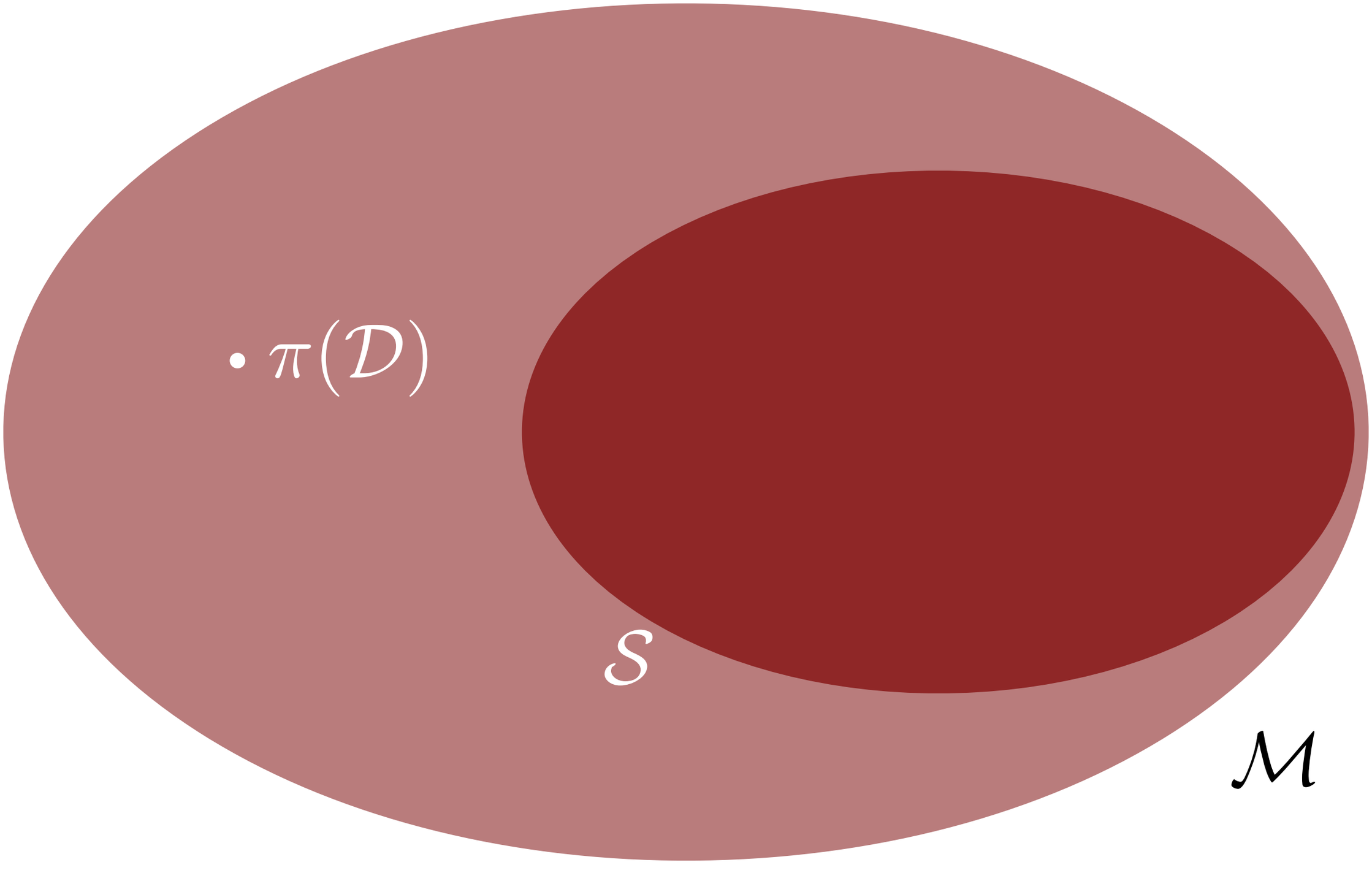
The "true" data generation process must lie in the space of all possible data generating processes, $\mathcal{M}$.



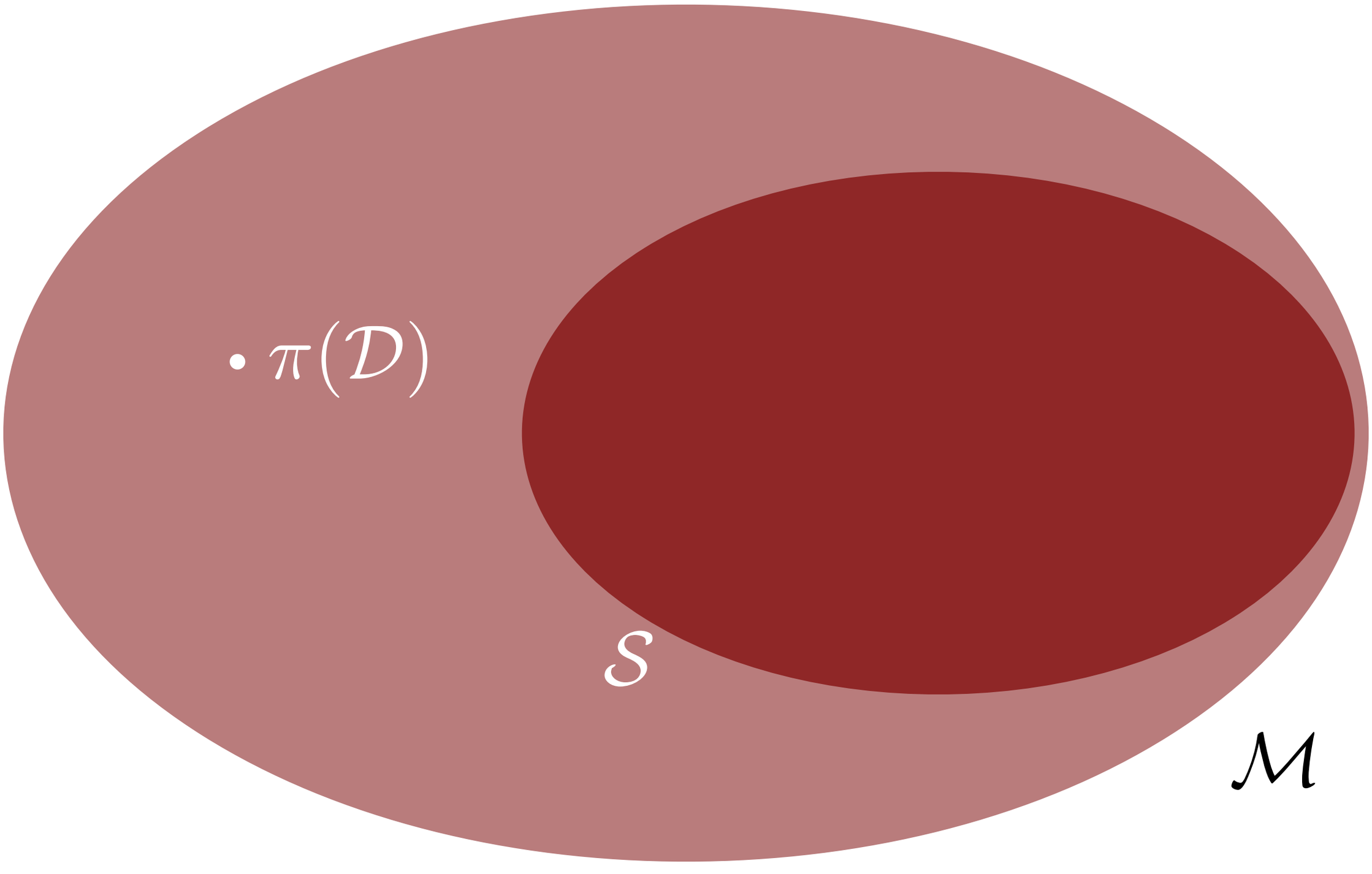
But in practice we have to consider only a small selection of processes $\mathcal{S} \subset \mathcal{M}$, sometimes called the *small world*.



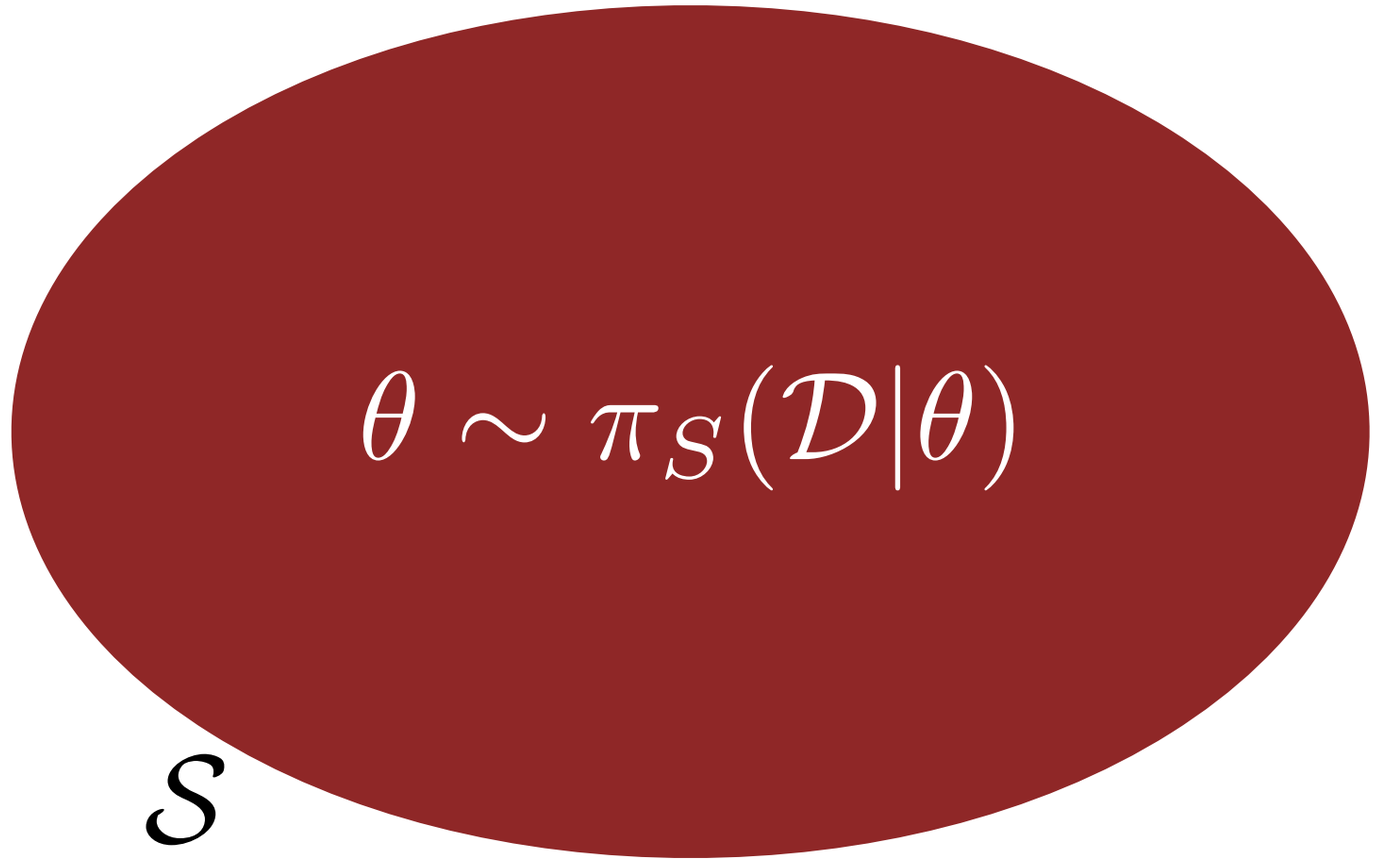
The true process, if it exists, may not be an element of the small world, but our inferences may still be meaningful.



"All models are wrong but some are useful".
-George Box



Any inferential model is then a choice of small world,
or a *likelihood* of distributions over measurements.



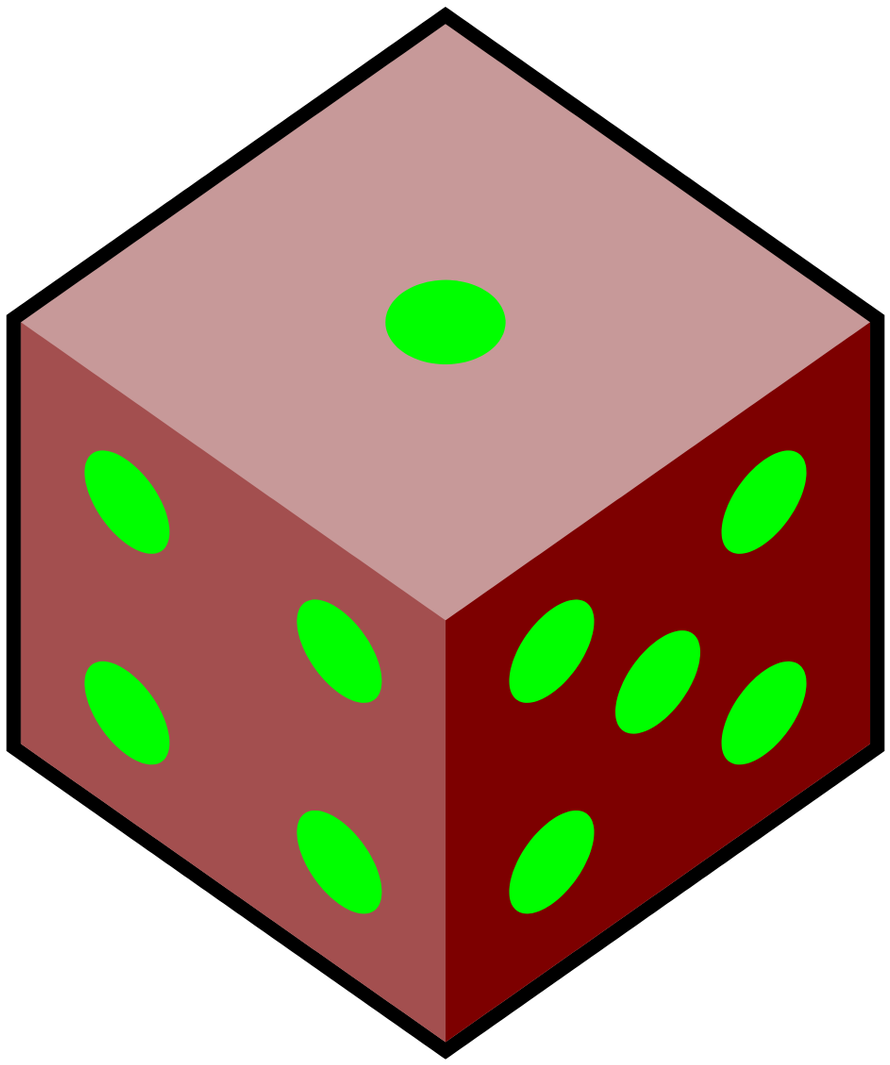
And inference is the identification of those points in the small world consistent with a given measurement.

$\widetilde{\mathcal{D}}$

$\theta \sim \pi_S(\mathcal{D}|\theta)?$

$\mathcal{S}$

How we define such an implementation, however, depends on how we define probability itself.



In *frequentist statistics*, probability is defined in terms of frequencies and so can be applied to only the data.

$$\pi_S(\mathcal{D}|\theta)$$

In *frequentist statistics*, probability is defined in terms of frequencies and so can be applied to only the data.

$$\pi_S(\mathcal{D}|\theta)$$

In *frequentist statistics*, probability is defined in terms of frequencies and so can be applied to only the data.

$$\pi_S(\mathcal{D}|\boldsymbol{\theta})$$

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*.

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

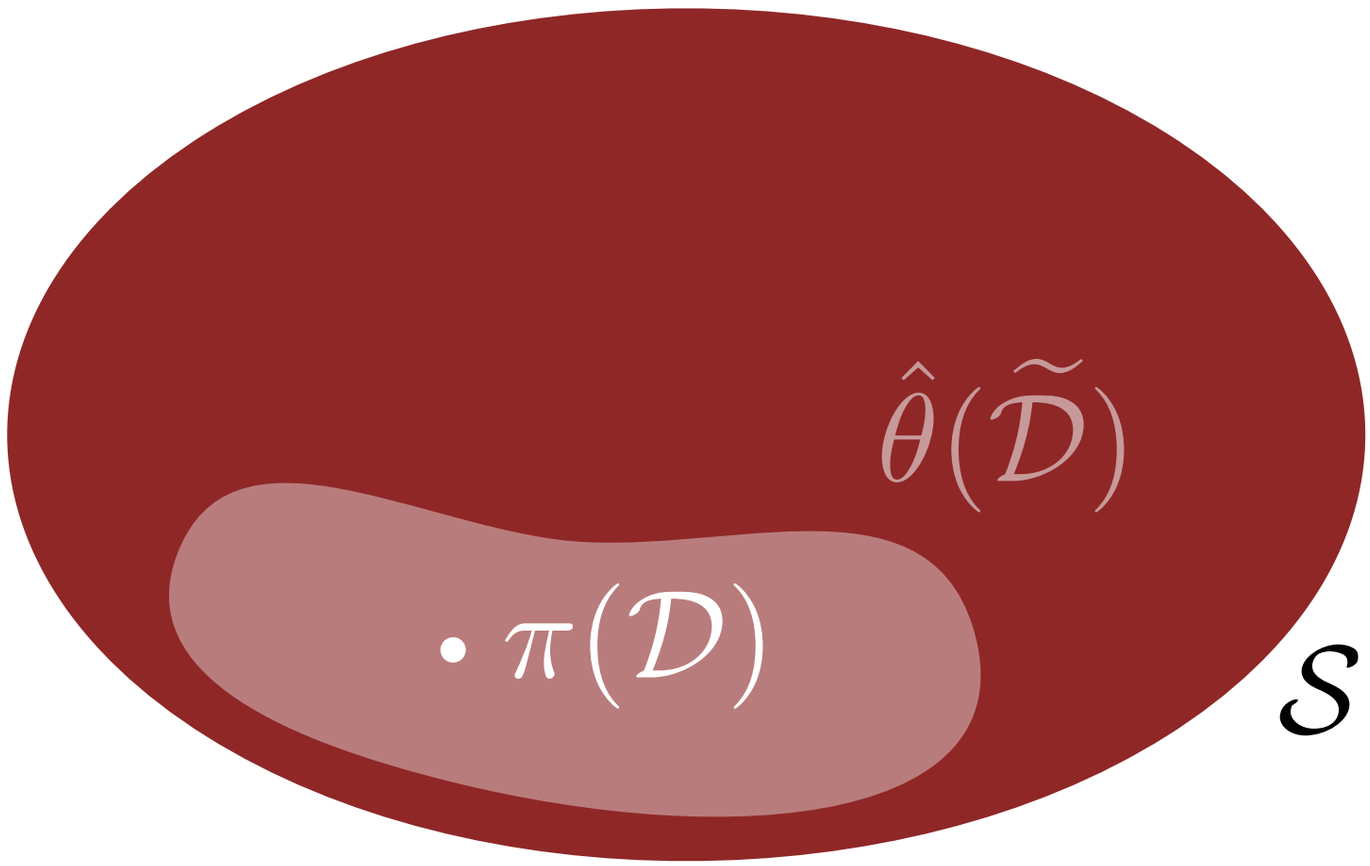
$$\mathcal{L}(\hat{\theta}(\mathcal{D}), \theta)$$

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

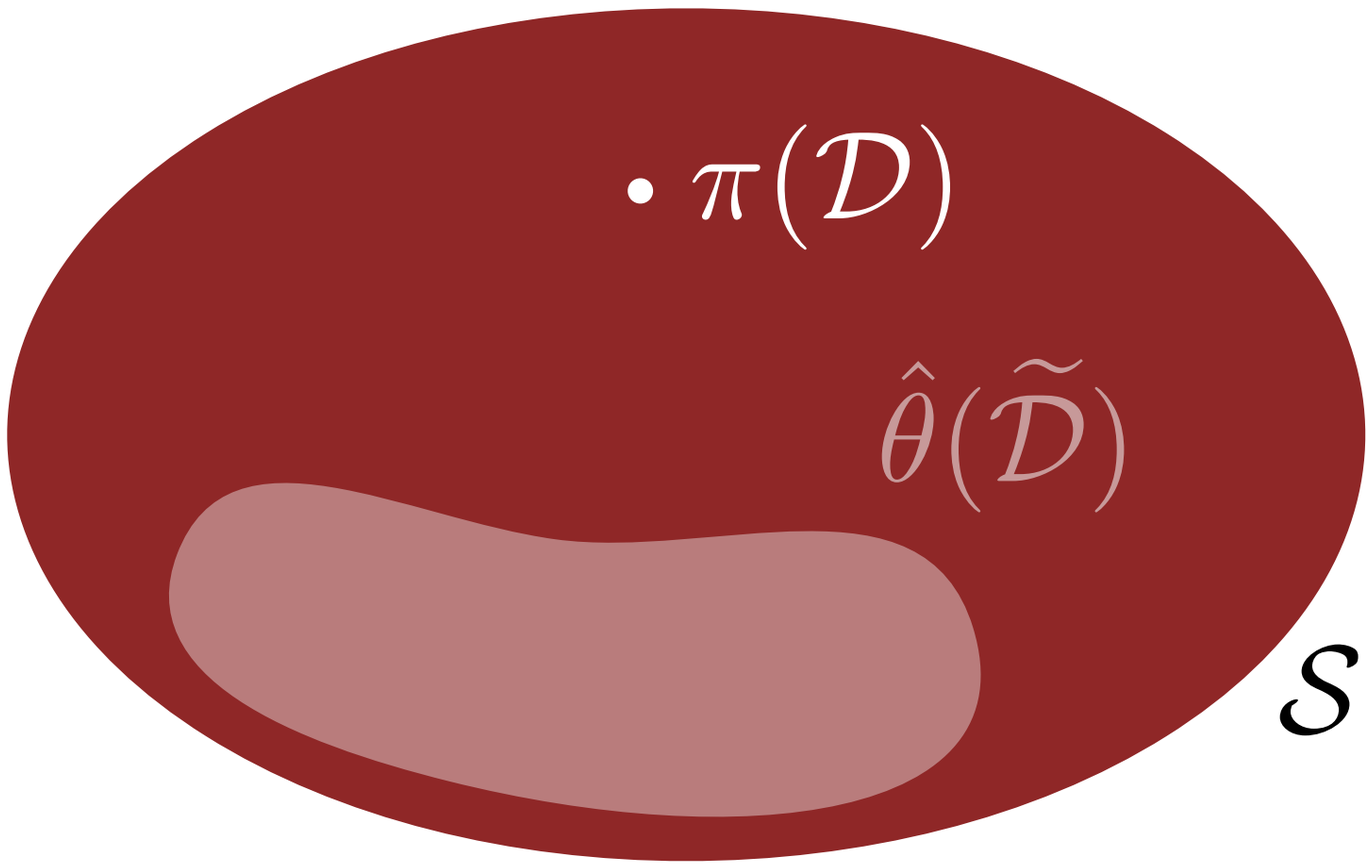
$$\hat{\theta}(\mathcal{D})$$

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

Frequentist methods must make hard decisions, selecting only a subset of small world for each measurement.



Frequentist methods must make hard decisions, selecting only a subset of small world for each measurement.



Bayesian inference, however, treats probability as a general measure of uncertainty.

Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi_S(\mathcal{D}|\theta)$$

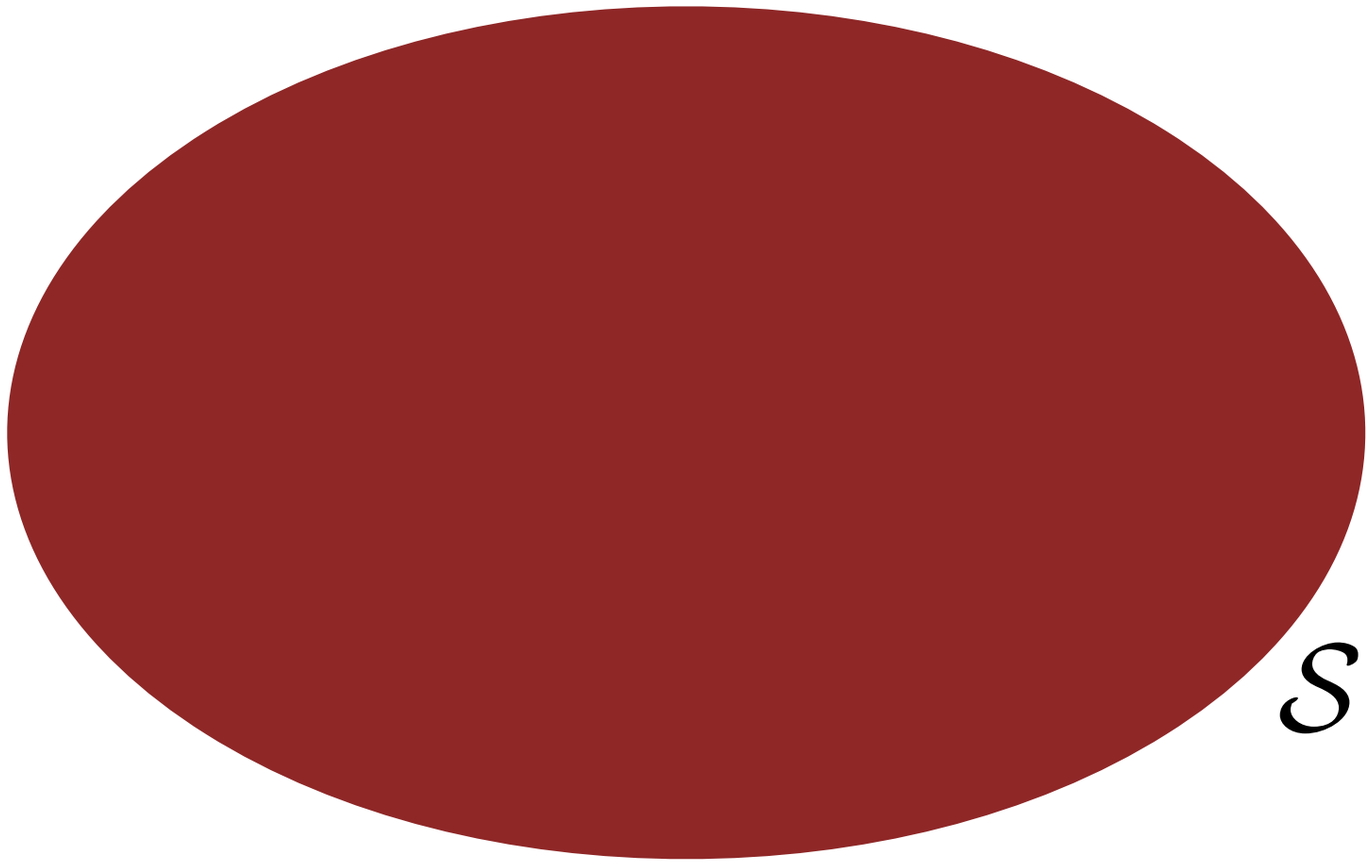
Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi_S(\mathcal{D}|\theta)$$

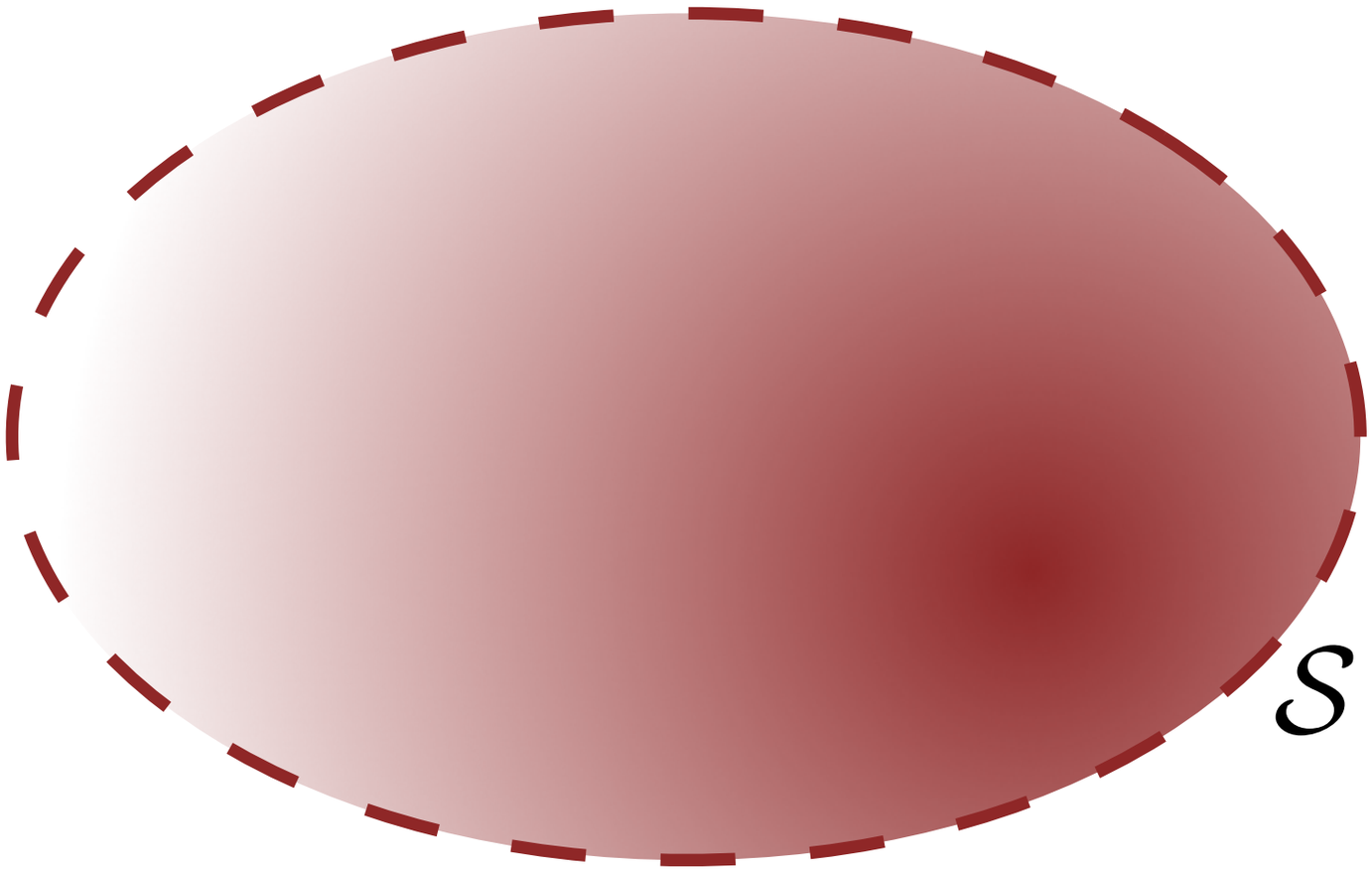
Bayesian inference builds upon frequentist inference by treating the data *and* the parameters as uncertain.

$$\pi_S(\mathcal{D}|\boldsymbol{\theta})$$

In this more general perspective we quantify consistency using a probability distribution over the small world.



In this more general perspective we quantify consistency using a probability distribution over the small world.



Uncertainty within the model is just
an application of *Bayes' Theorem*.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

The *prior* incorporates any prior knowledge about the model space before the data are measured.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

The *likelihood* is similar to the frequentist approach:
a generative model of the data.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

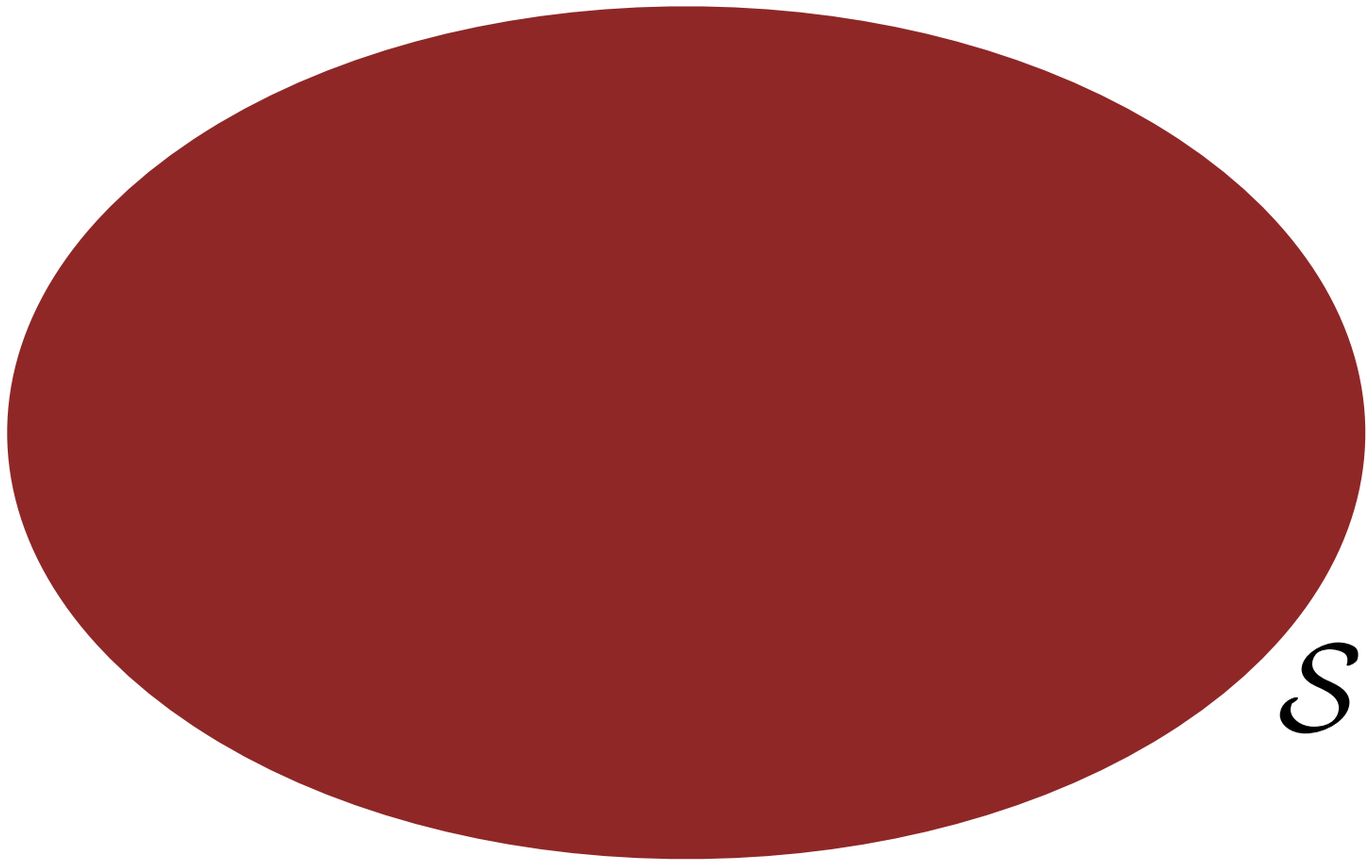
The *posterior* incorporates both the information into the prior and the data into a final uncertainty in the model.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

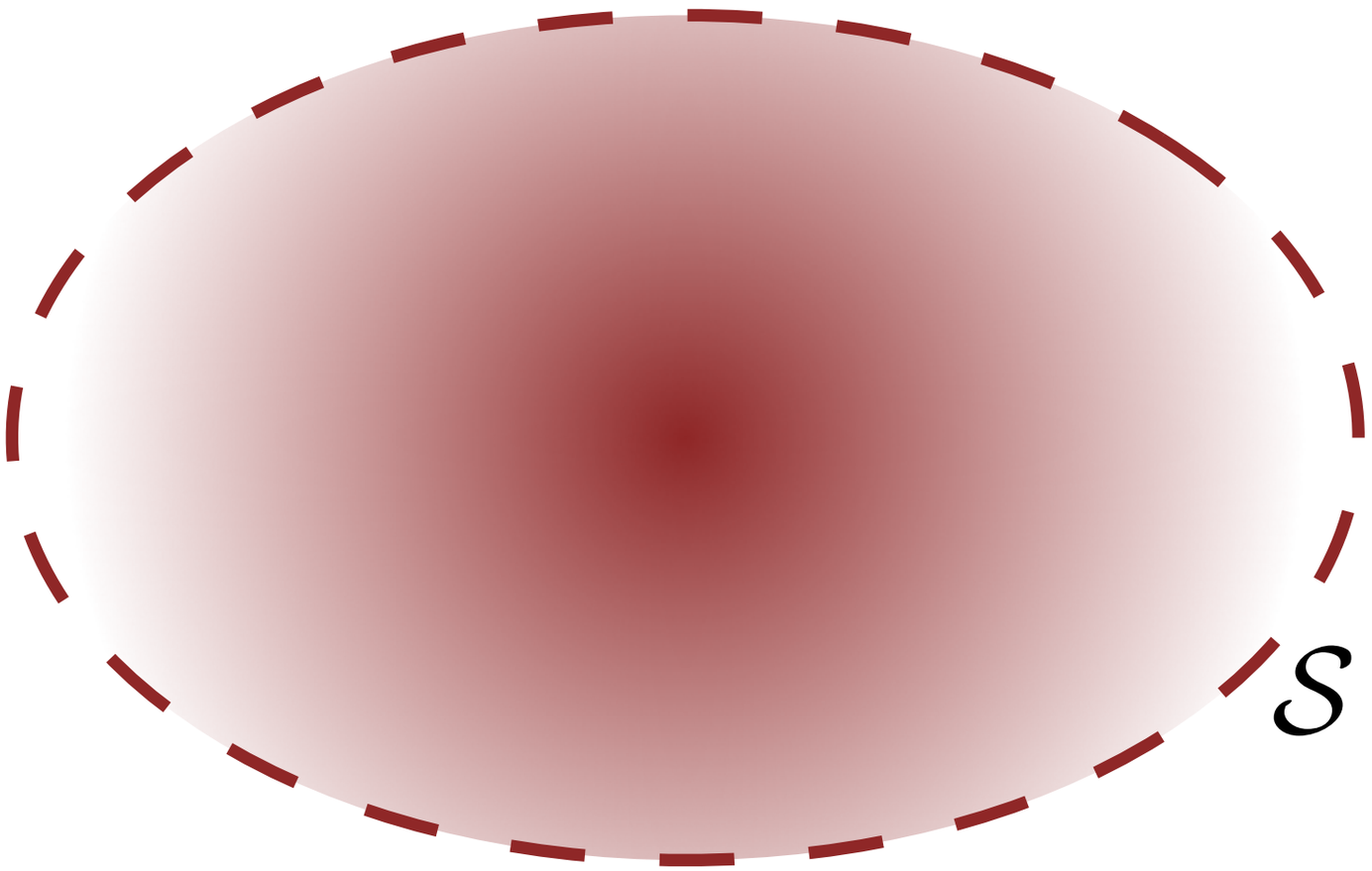
The *marginal likelihood* gives the probability of the measurement conditioned on the model.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)}{\pi_S(\widetilde{\mathcal{D}})}$$

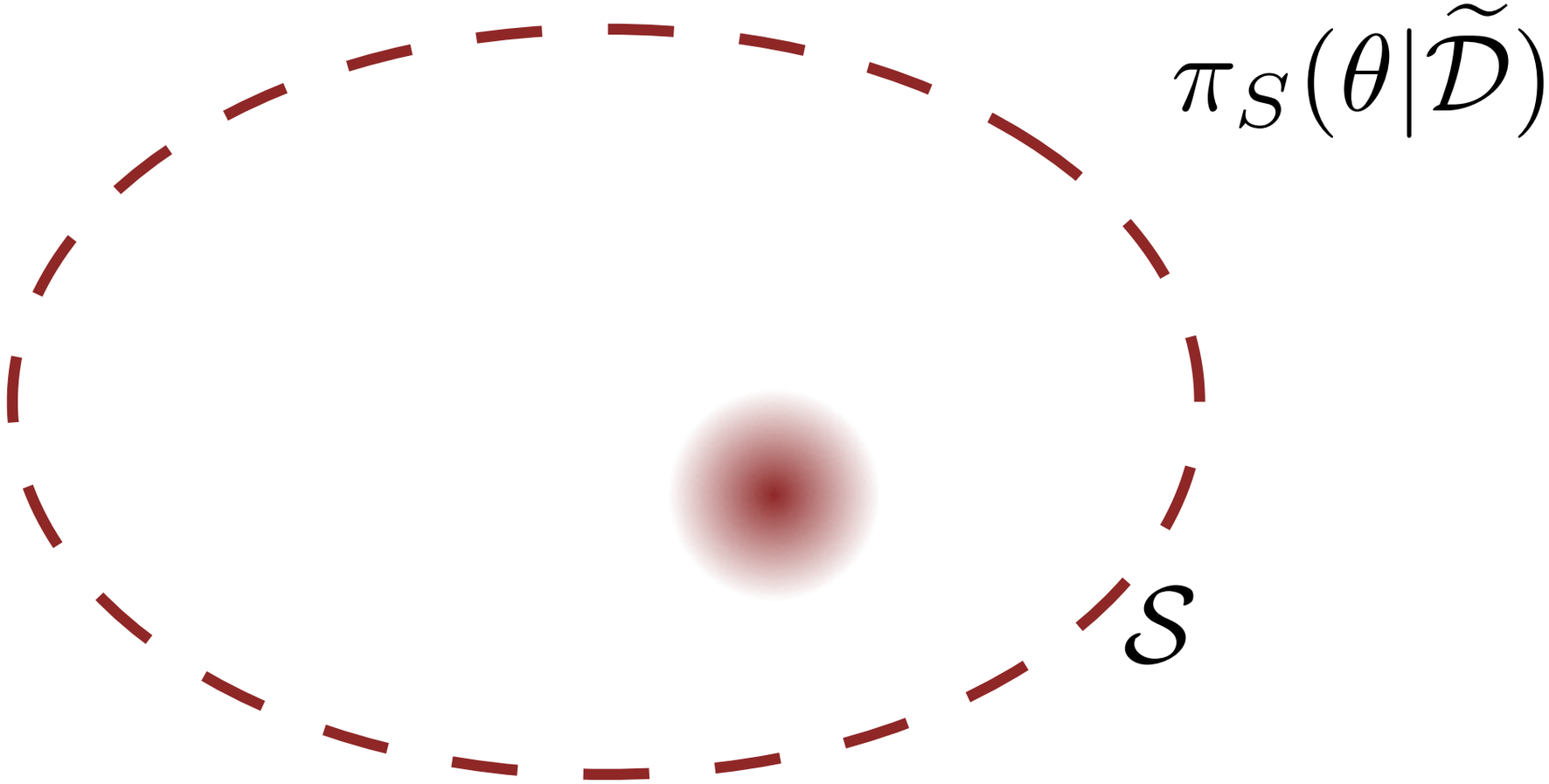
Conditioning on the measurement reduces our uncertainty amongst the data generation processes.



Conditioning on the measurement reduces our uncertainty amongst the data generation processes.



Conditioning on the measurement reduces our uncertainty amongst the data generation processes.



From a Bayesian perspective, all inferential questions are answered by expectations.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta)$$

Expectations include means and variances for posterior summaries and expected utility for decision making.

$$\mu = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta$$

Expectations include means and variances for posterior summaries and expected utility for decision making.

$$\mu = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta$$

$$\sigma^2 = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta^2 - \mu^2$$

Expectations include means and variances for posterior summaries and expected utility for decision making.

$$\mu = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta$$

$$\sigma^2 = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, \theta^2 - \mu^2$$

$$U(A) = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, U(A, \theta)$$

Anything sensitive to a particular parameterization, such as the posterior mode, however, is *not* a valid expectation!

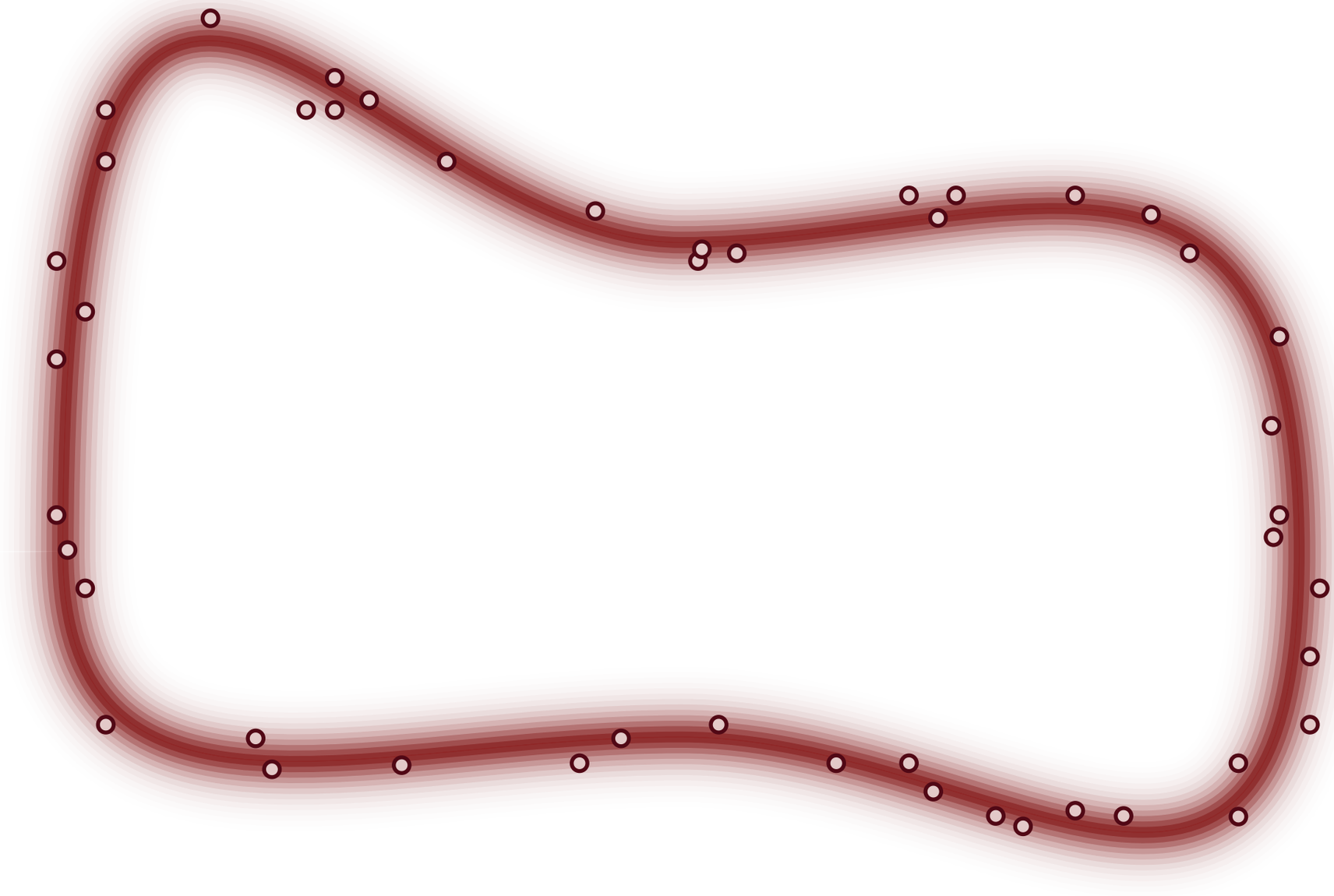
$$\hat{\theta} = \underset{\theta}{\mathrm{argmax}}\, \pi_S(\theta|\widetilde{\mathcal{D}})$$

Anything sensitive to a particular parameterization, such as the posterior mode, however, is *not* a valid expectation!

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, \pi_S(\theta|\widetilde{\mathcal{D}})$$

$$\pi_S(\phi(\theta)\,|\widetilde{\mathcal{D}}) = \pi_S(\theta|\widetilde{\mathcal{D}}) \left| \frac{\partial \phi}{\partial \theta} \right|$$

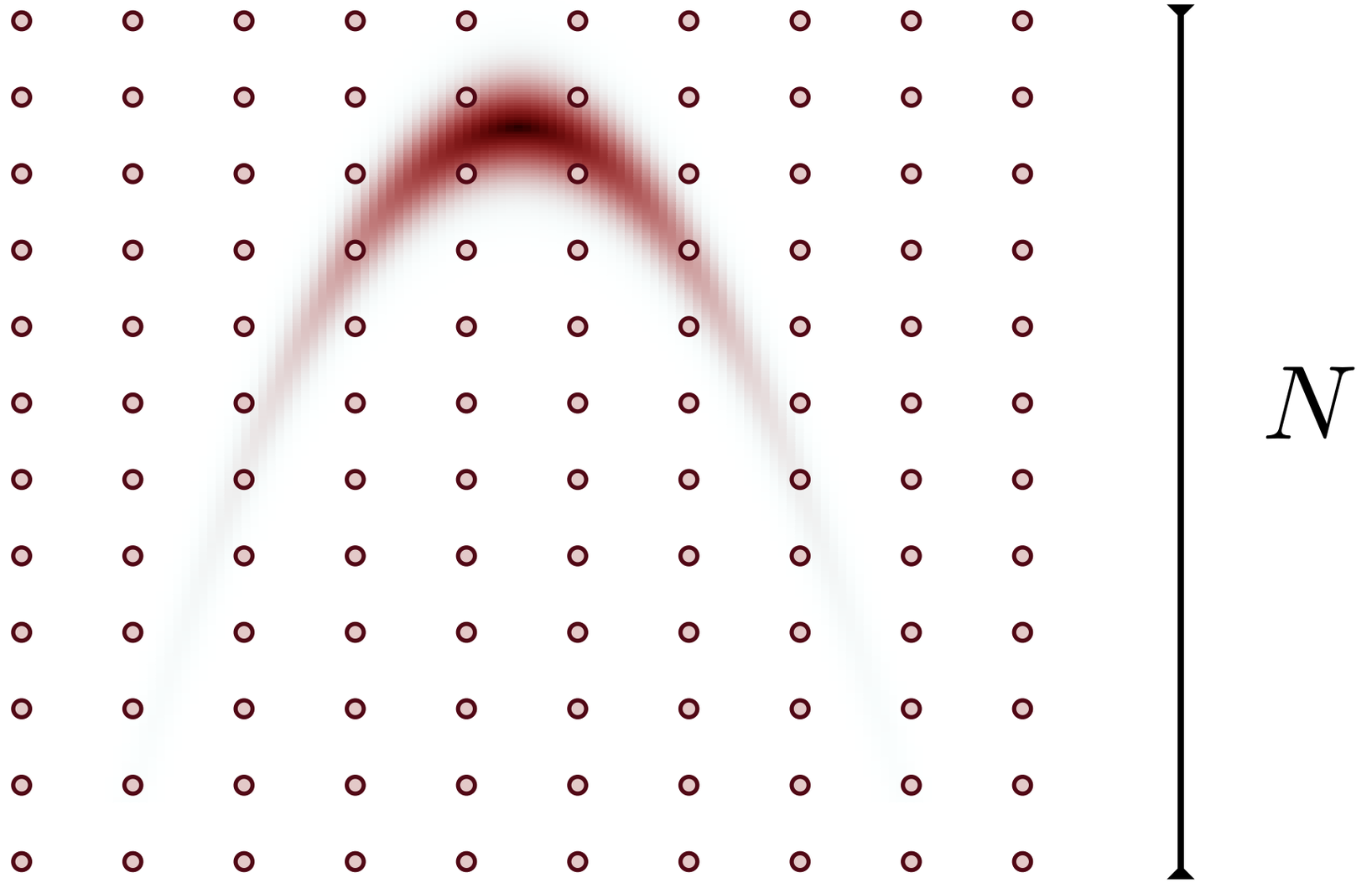
Foundations of Bayesian Computation



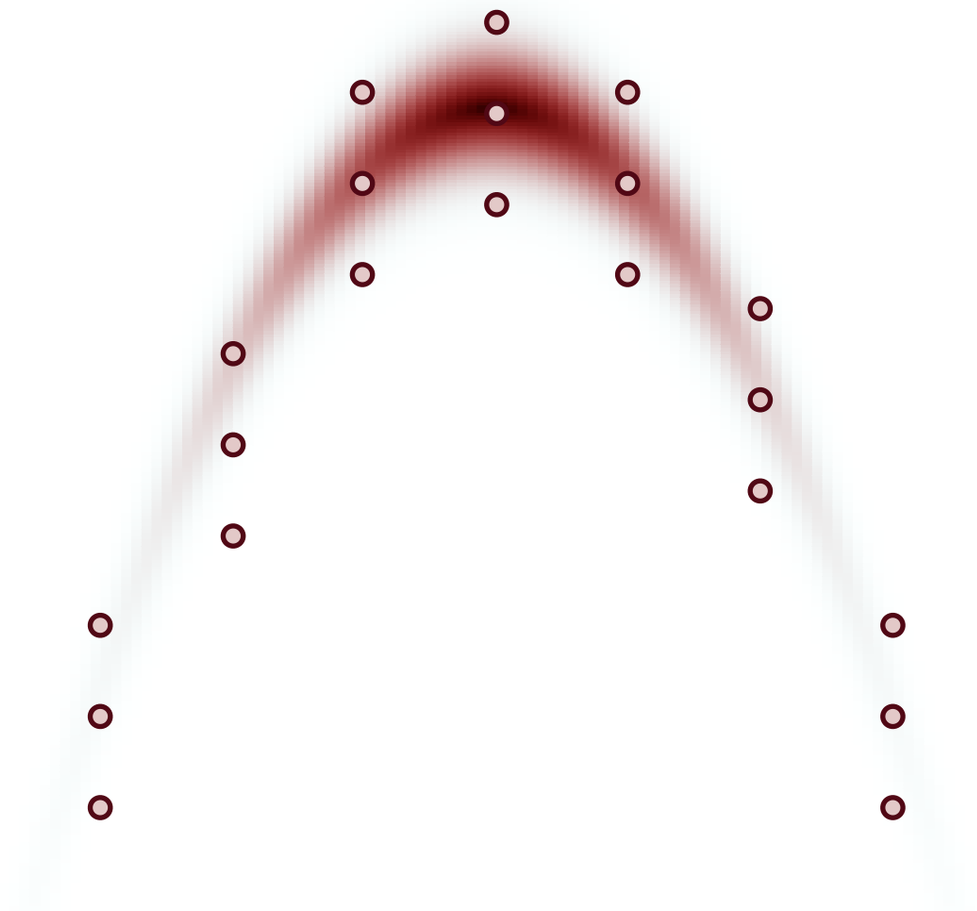
Once we've built a model, Bayesian computation reduces to evaluating expectations, or integrals.

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta)$$

Unfortunately, the cost of naive algorithms scales exponentially with the dimension of the posterior.



To be efficient we need to focus computation on the relevant neighborhoods of parameter space.



But exactly which neighborhoods end up contributing most to arbitrary expectations?

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta)$$

But exactly which neighborhoods end up contributing most to arbitrary expectations?

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, f(\theta)$$

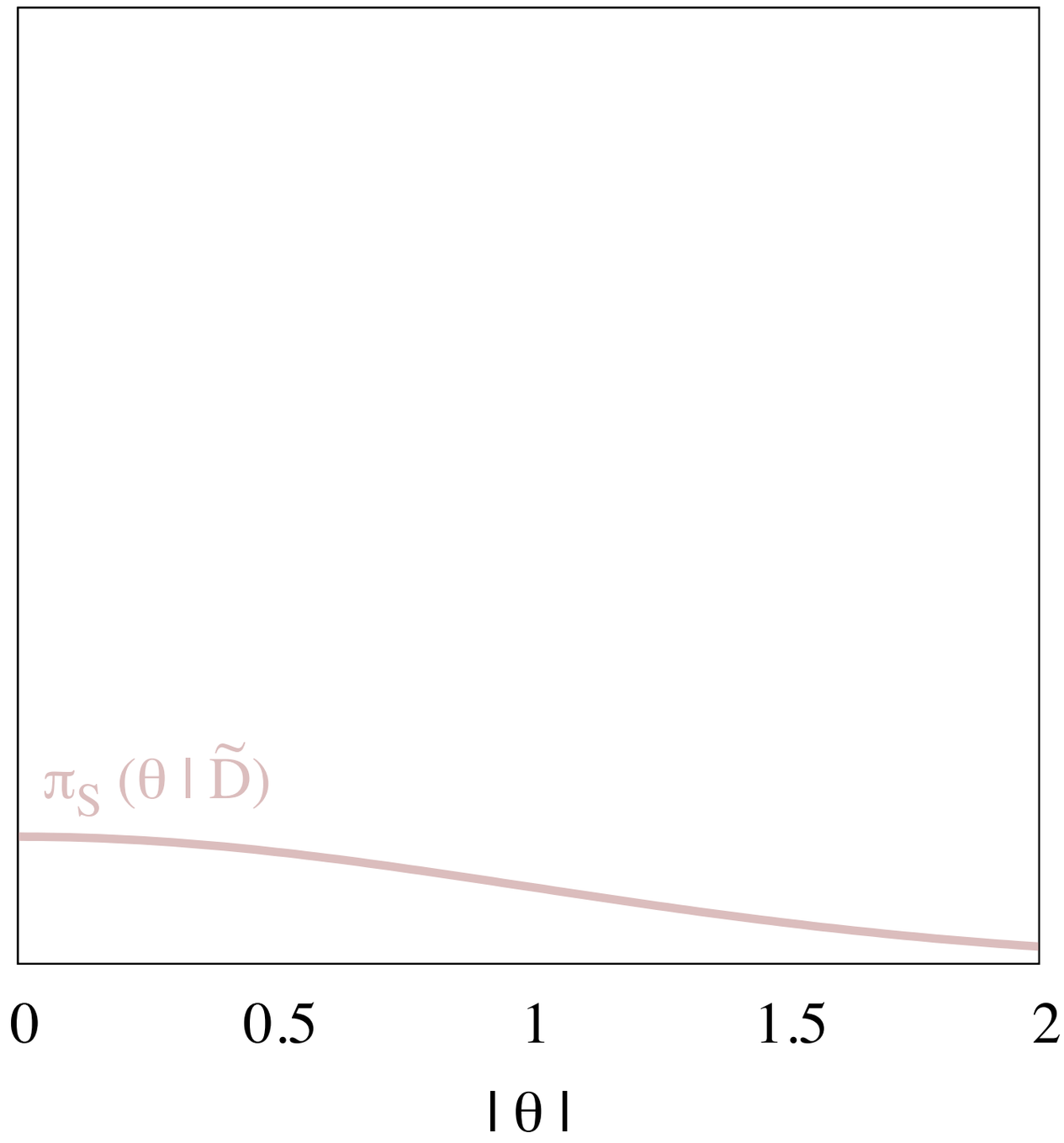
But exactly which neighborhoods end up contributing most to arbitrary expectations?

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) \, f(\theta)$$

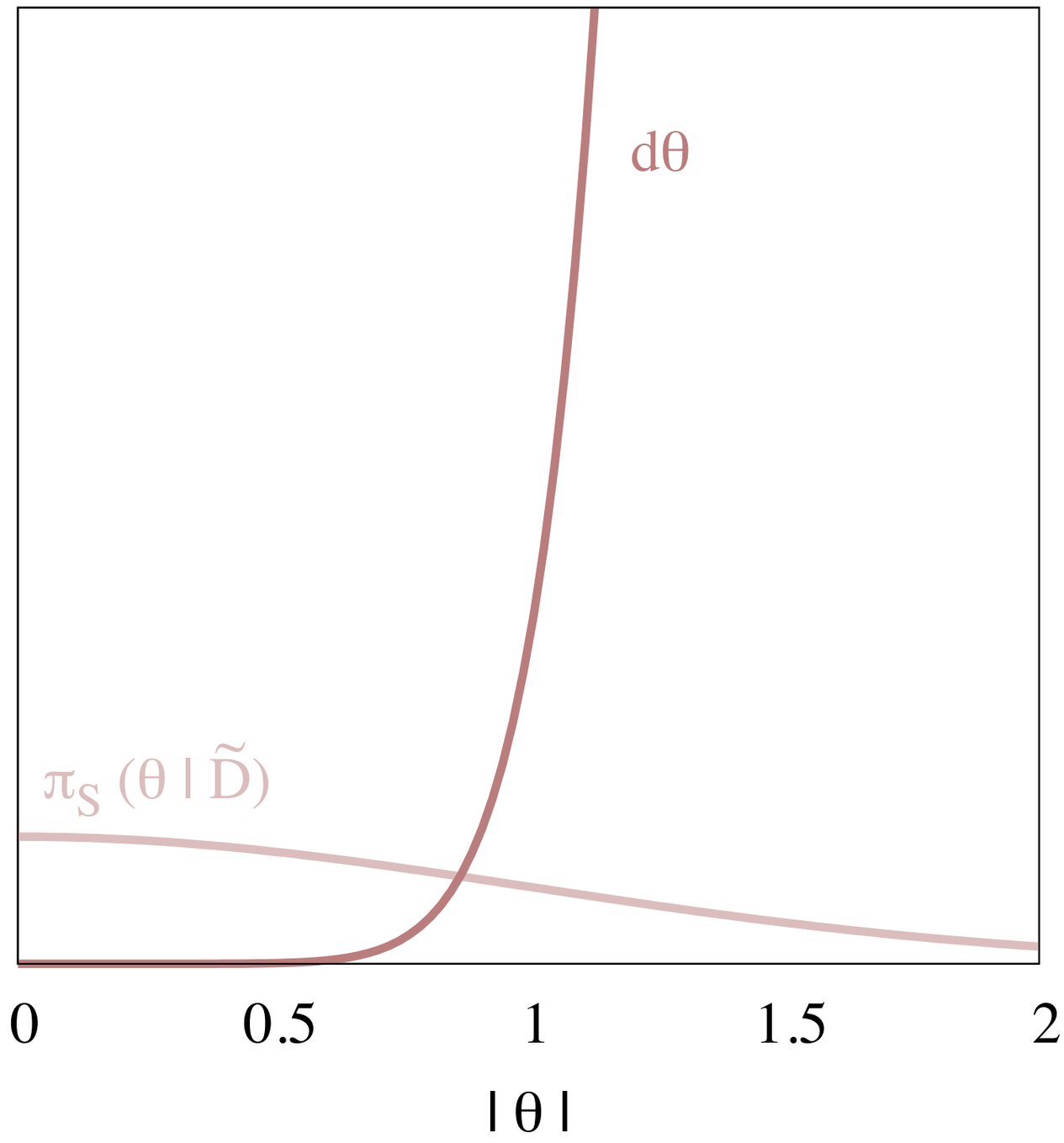
But exactly which neighborhoods end up contributing most to arbitrary expectations?

$$\mathbb{E}[f] = \int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta)$$

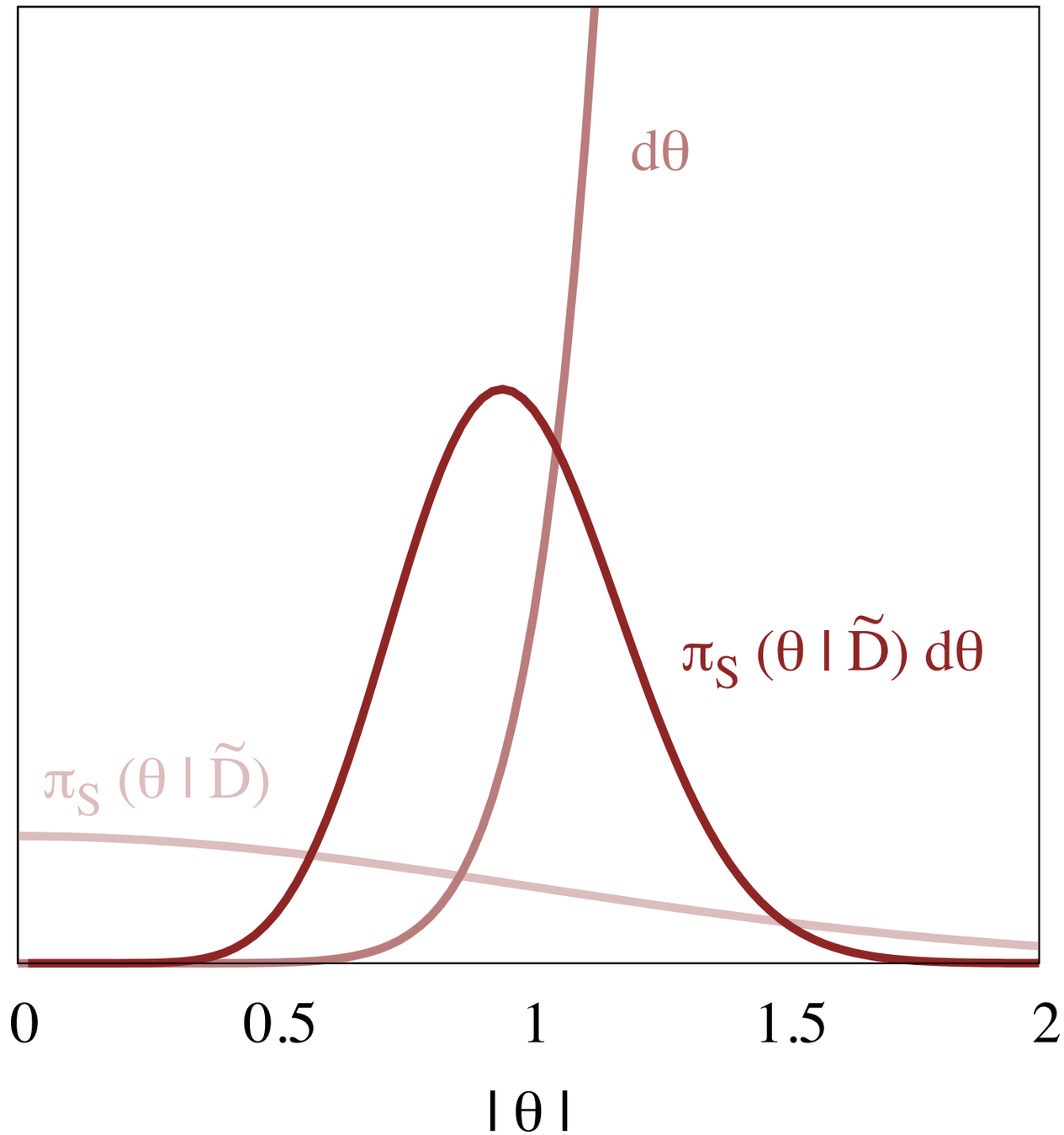
Relevant neighborhoods, however, are defined not by probability density but rather by probability mass.



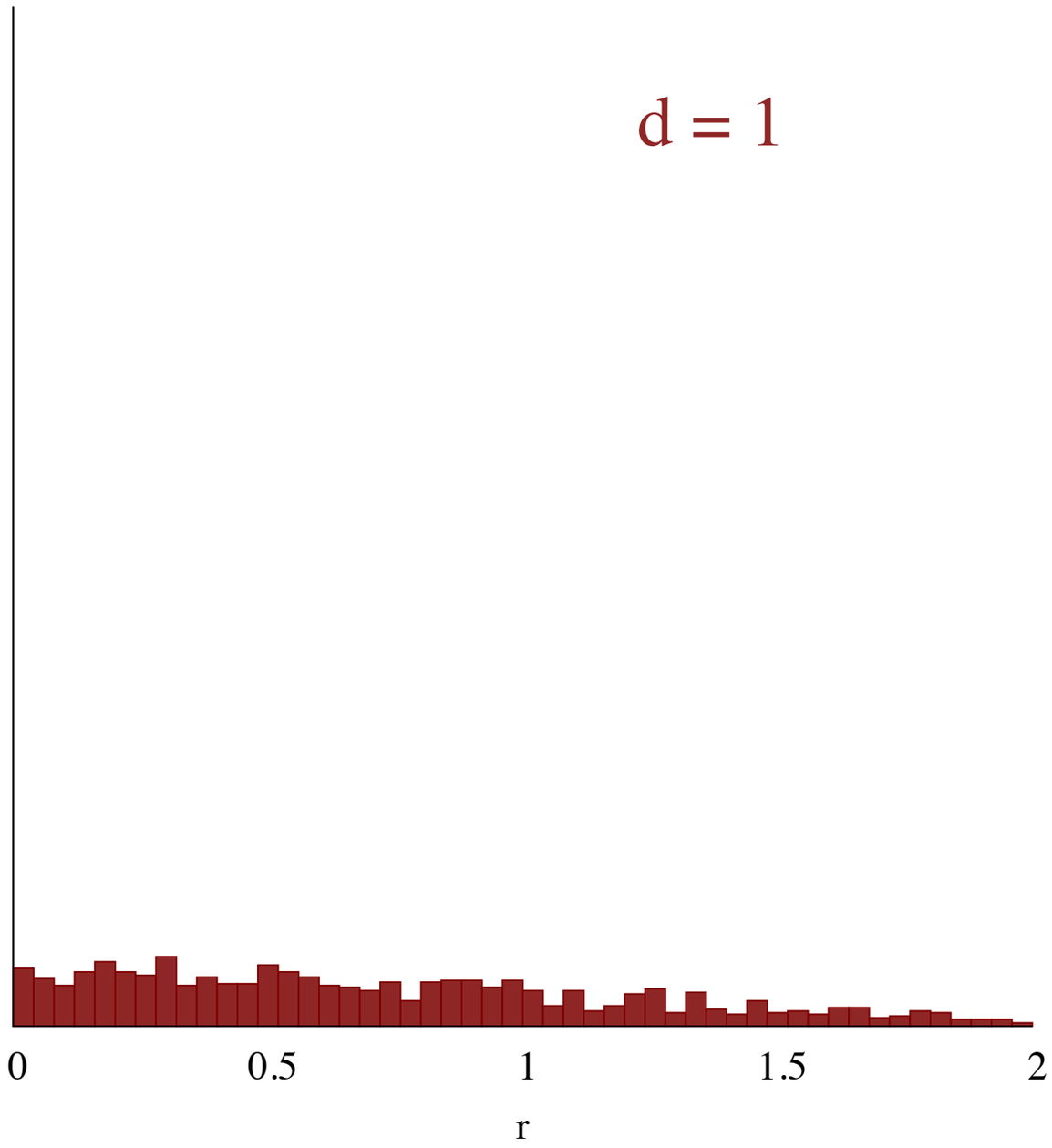
Relevant neighborhoods, however, are defined not by probability density but rather by probability mass.



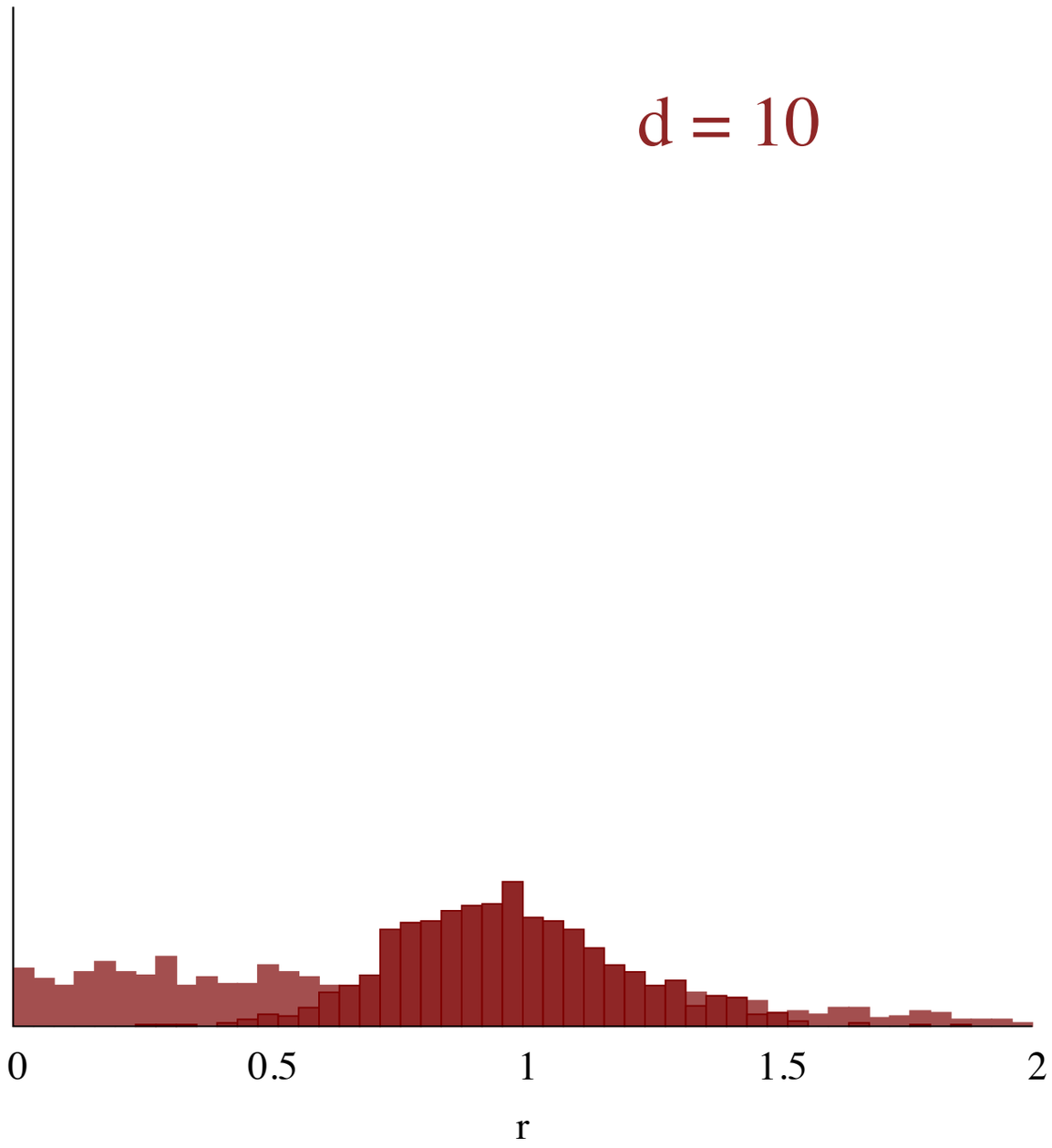
Relevant neighborhoods, however, are defined not by probability density but rather by probability mass.



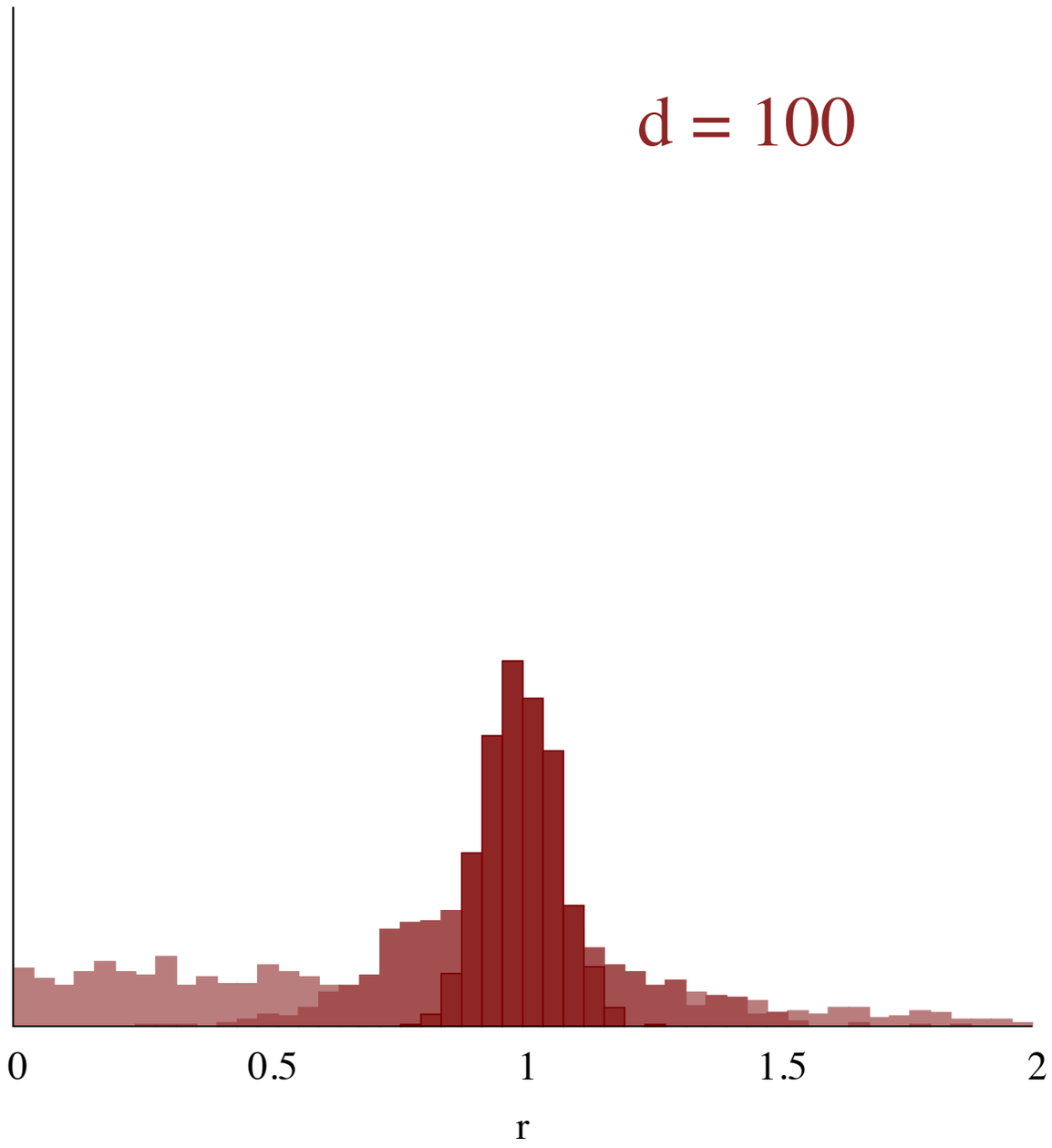
Probability mass concentrates on a hypersurface called the *typical set* that surrounds the mode.



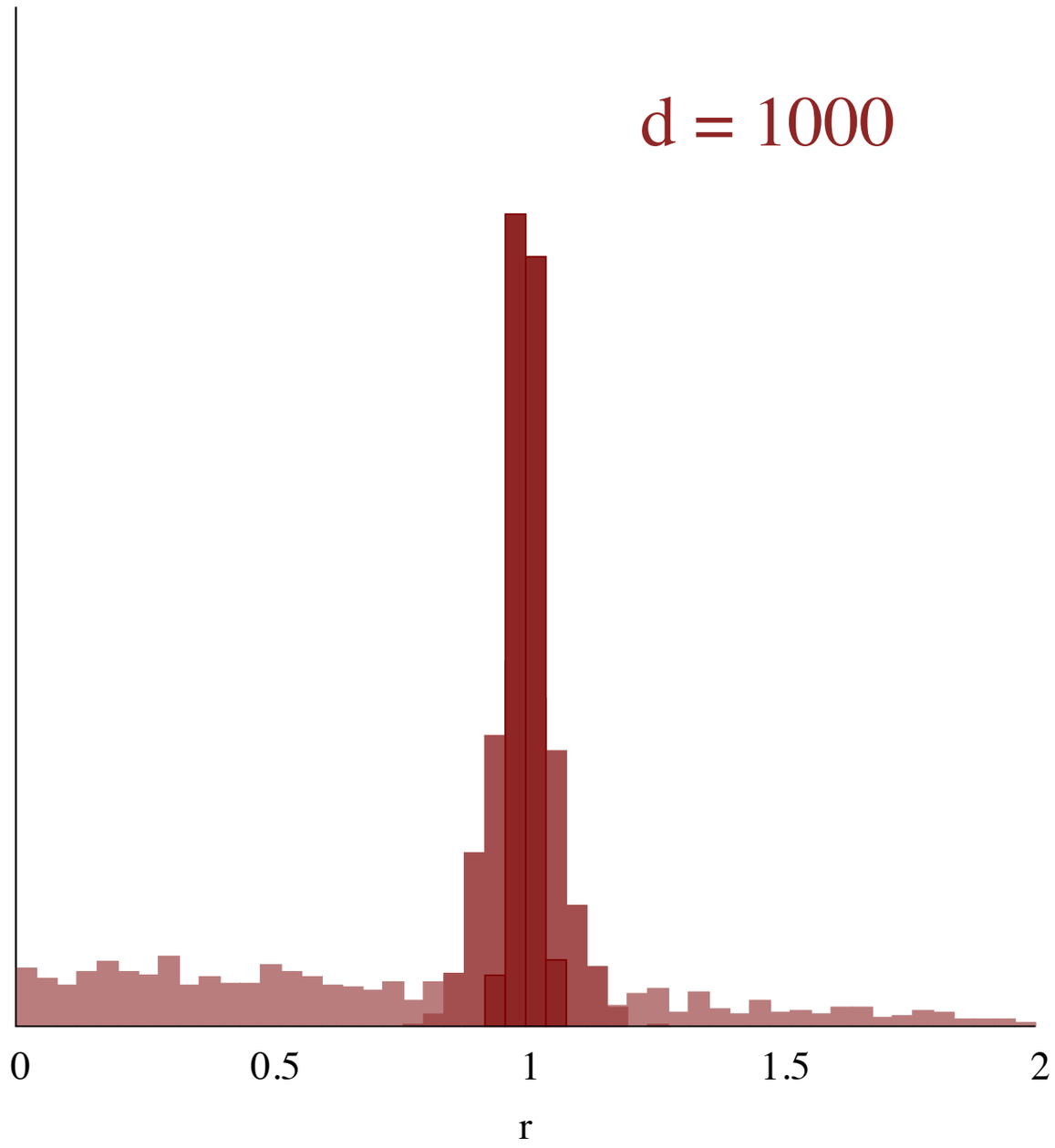
Probability mass concentrates on a hypersurface called the *typical set* that surrounds the mode.



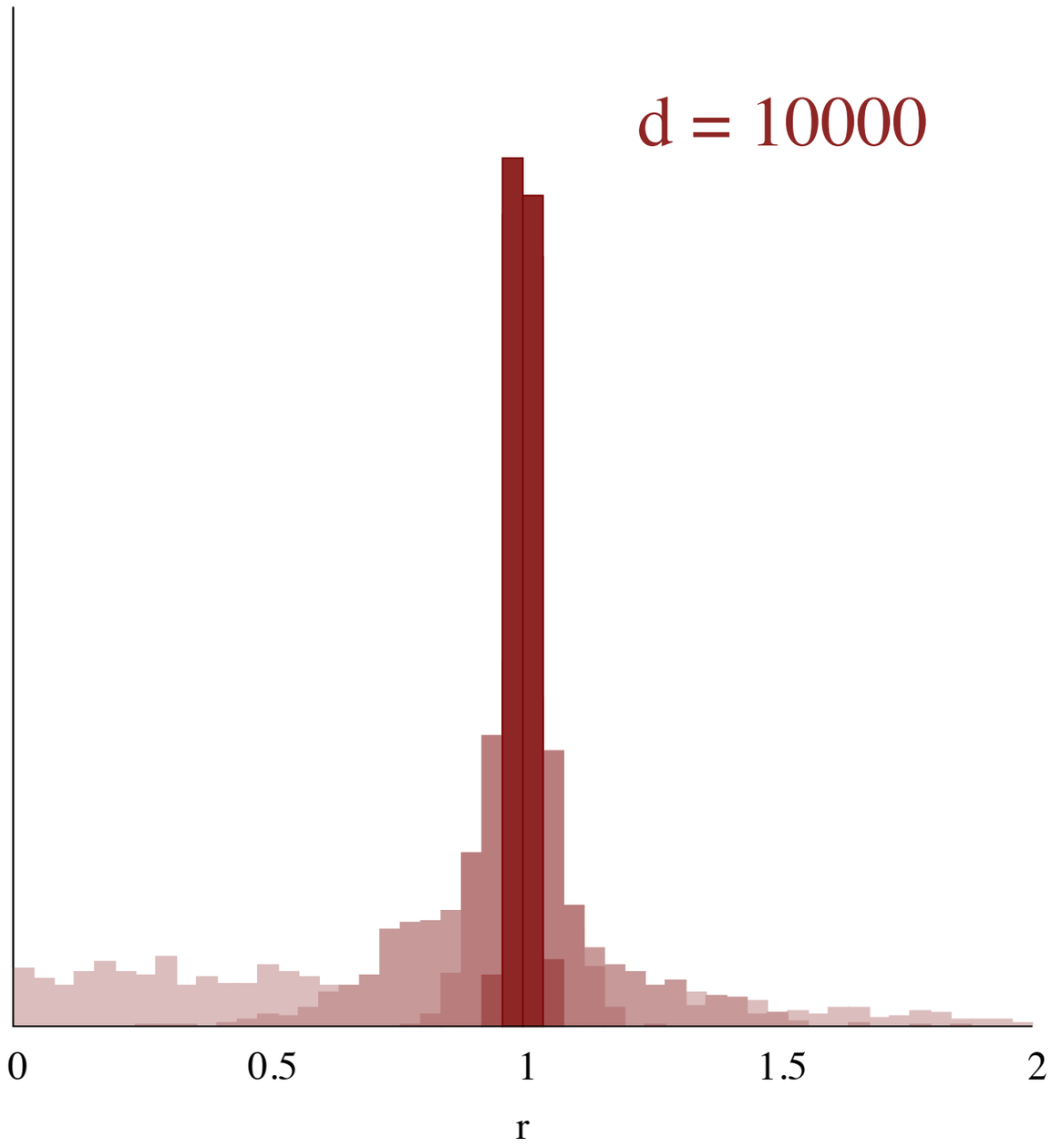
Probability mass concentrates on a hypersurface called the *typical set* that surrounds the mode.



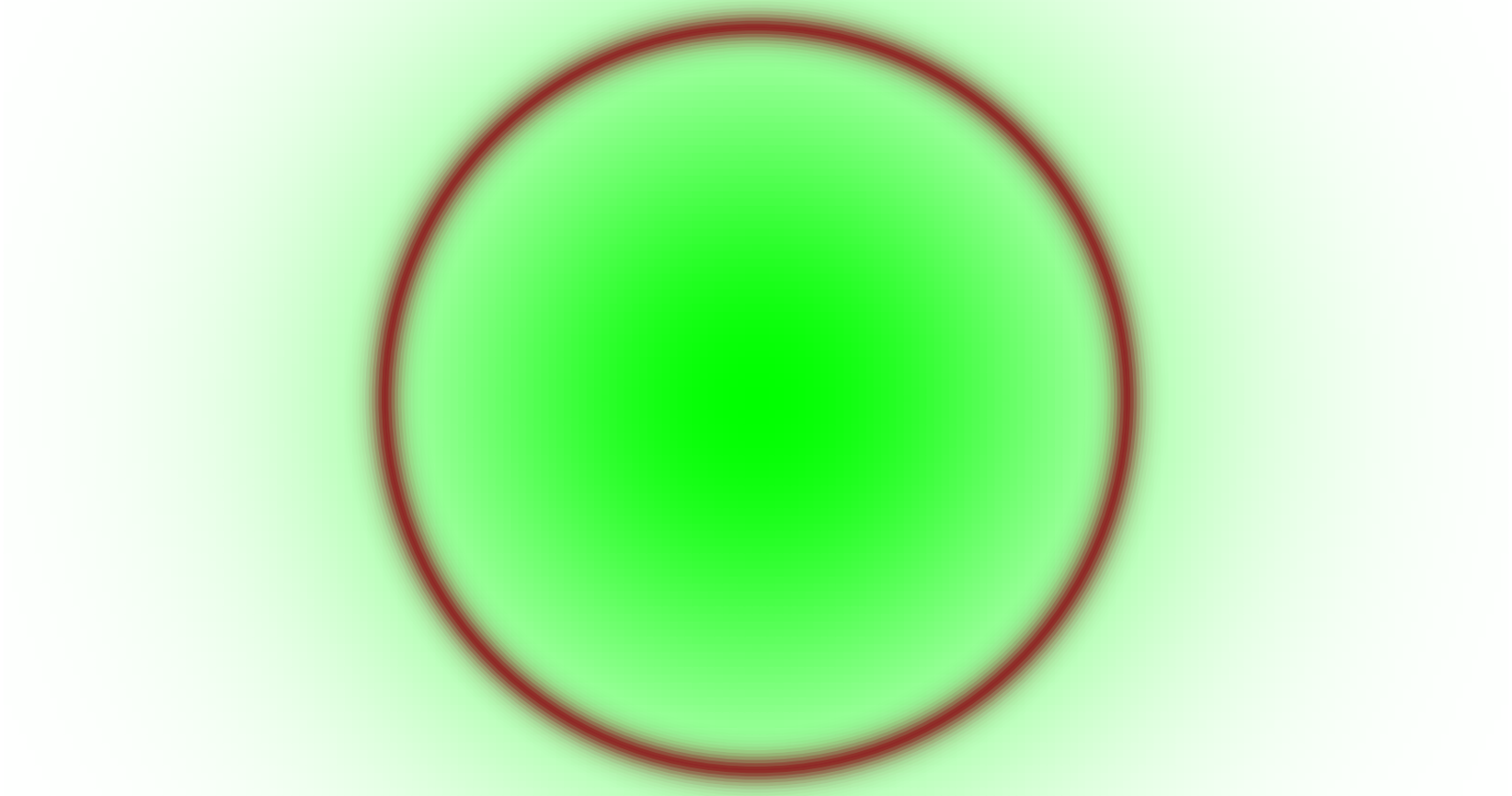
Probability mass concentrates on a hypersurface called the *typical set* that surrounds the mode.



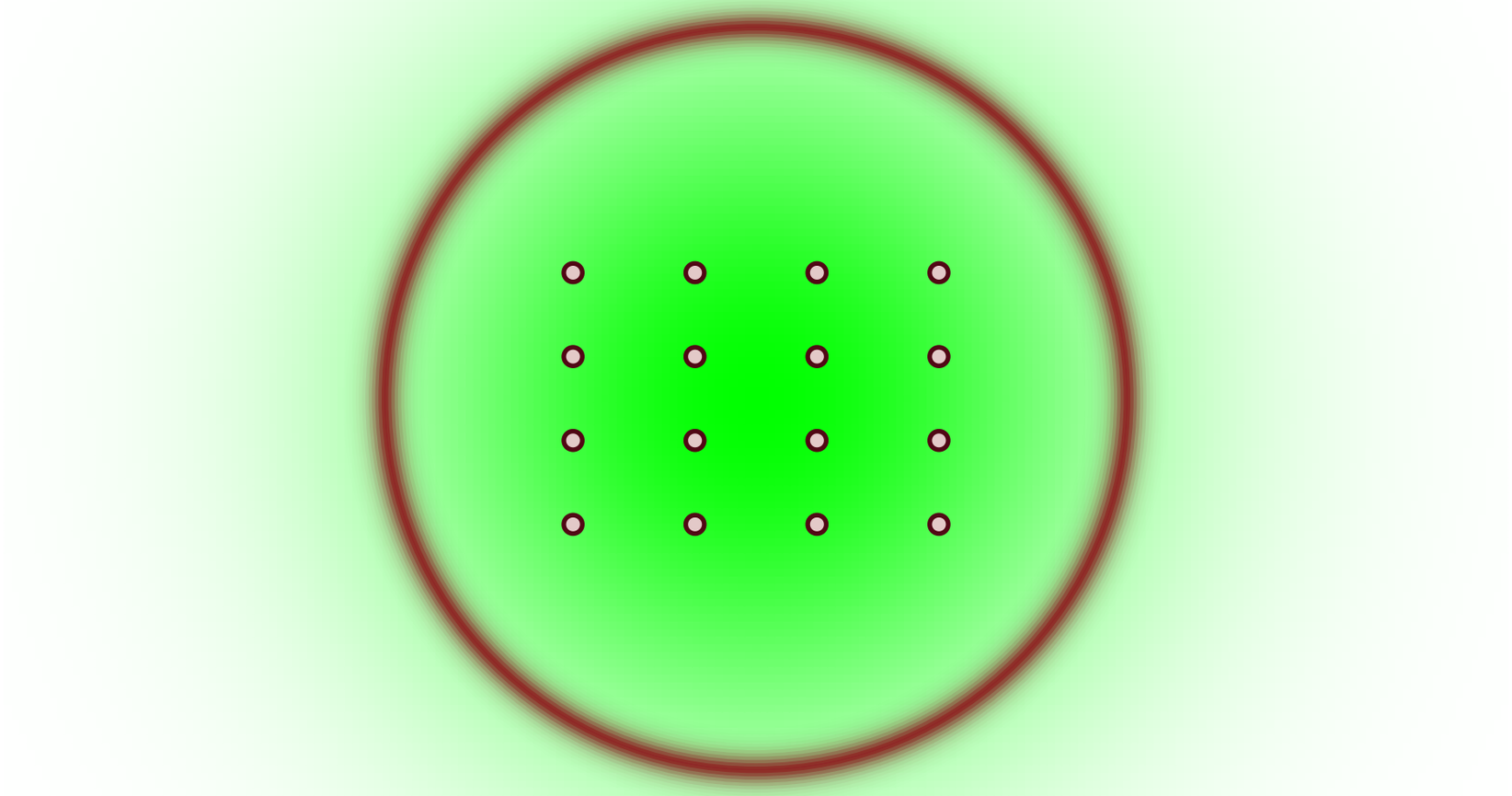
Probability mass concentrates on a hypersurface called the *typical set* that surrounds the mode.



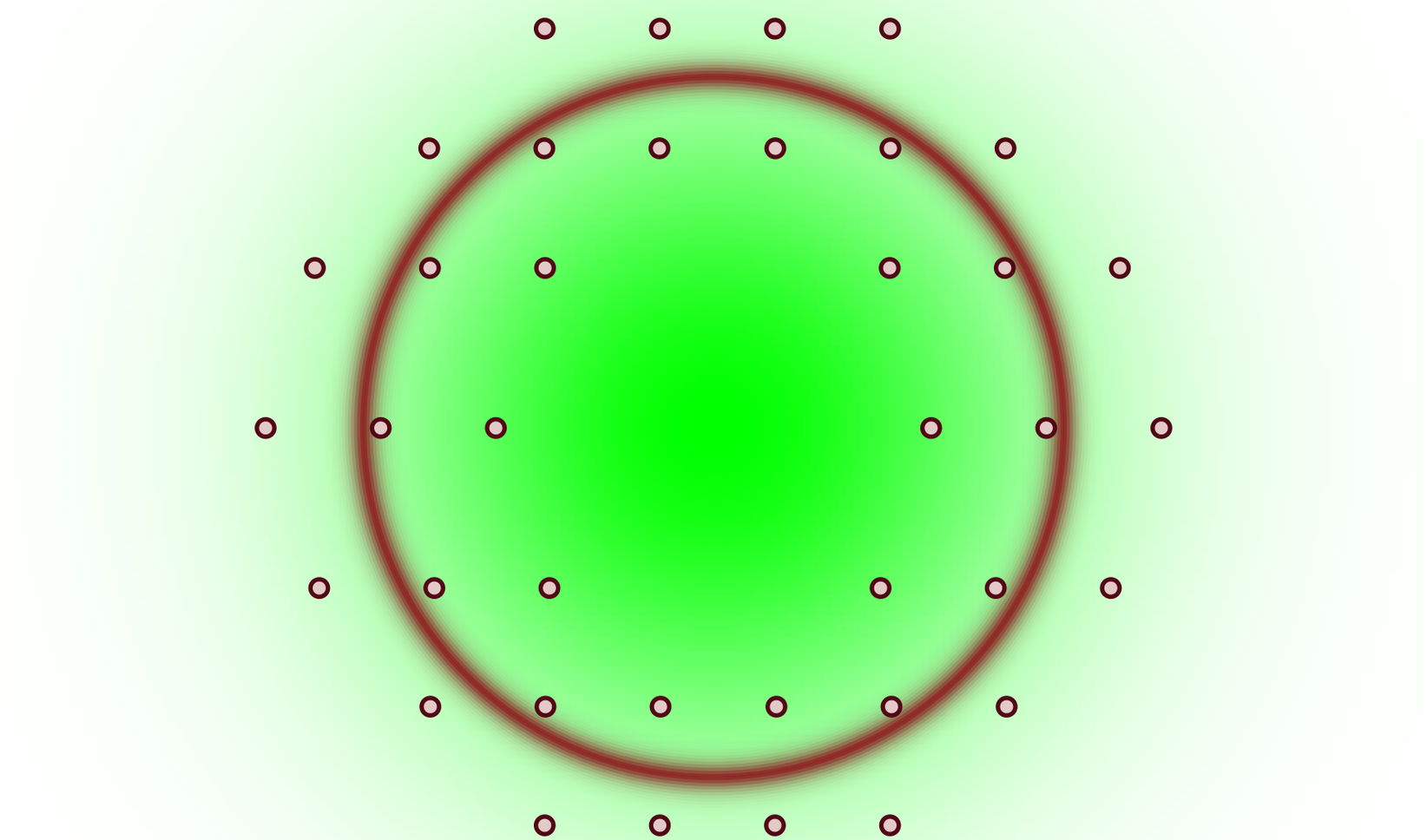
Identifying the typical set is analytically intractable,
even for seemingly simple distributions.



Identifying the typical set is analytically intractable,
even for seemingly simple distributions.



Identifying the typical set is analytically intractable,
even for seemingly simple distributions.



Of course, the problem becomes substantially more difficult for the complex distributions in practice.

In order to implement Bayesian inference in practice
we need better ways of approximating expectations.

Deterministic

Modal Estimators

Laplace Estimators

Variational Estimators

...

Stochastic

Rejection Sampling

Importance Sampling

Markov Chain Monte Carlo

...

Deterministic methods approximate posterior expectations with those from a simpler distribution.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) \approx q(\theta)$$

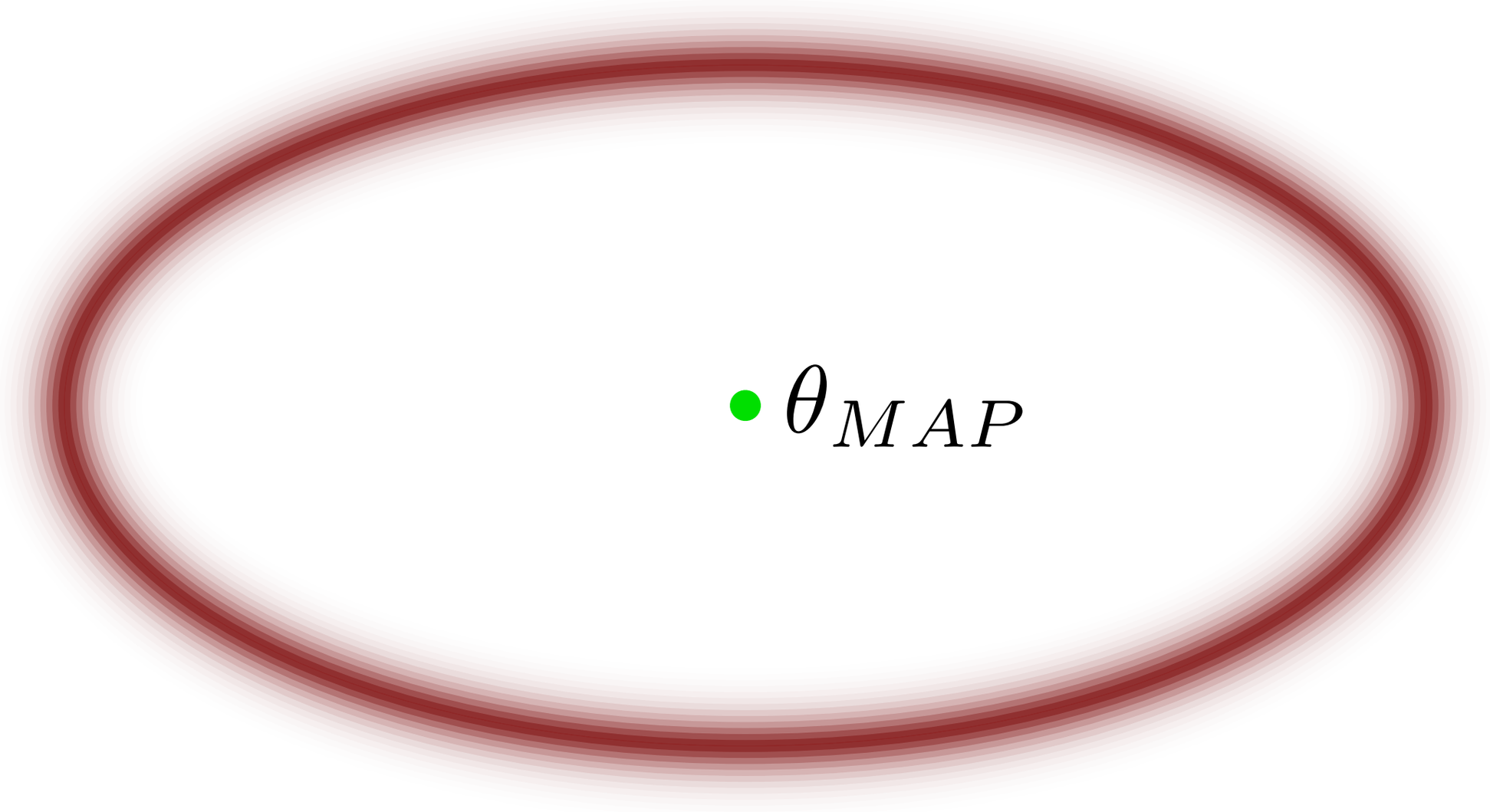
$$\int \mathrm{d}\theta\, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta) \approx \int \mathrm{d}\theta\, q(\theta) f(\theta)$$

MAP estimators approximate expectations with point estimates at a posterior density mode.

$$\pi_S(\theta|\widetilde{\mathcal{D}}) \approx \delta(\theta_{MAP})$$

$$\int \mathrm{d}\theta\, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta) \approx f(\theta_{MAP})$$

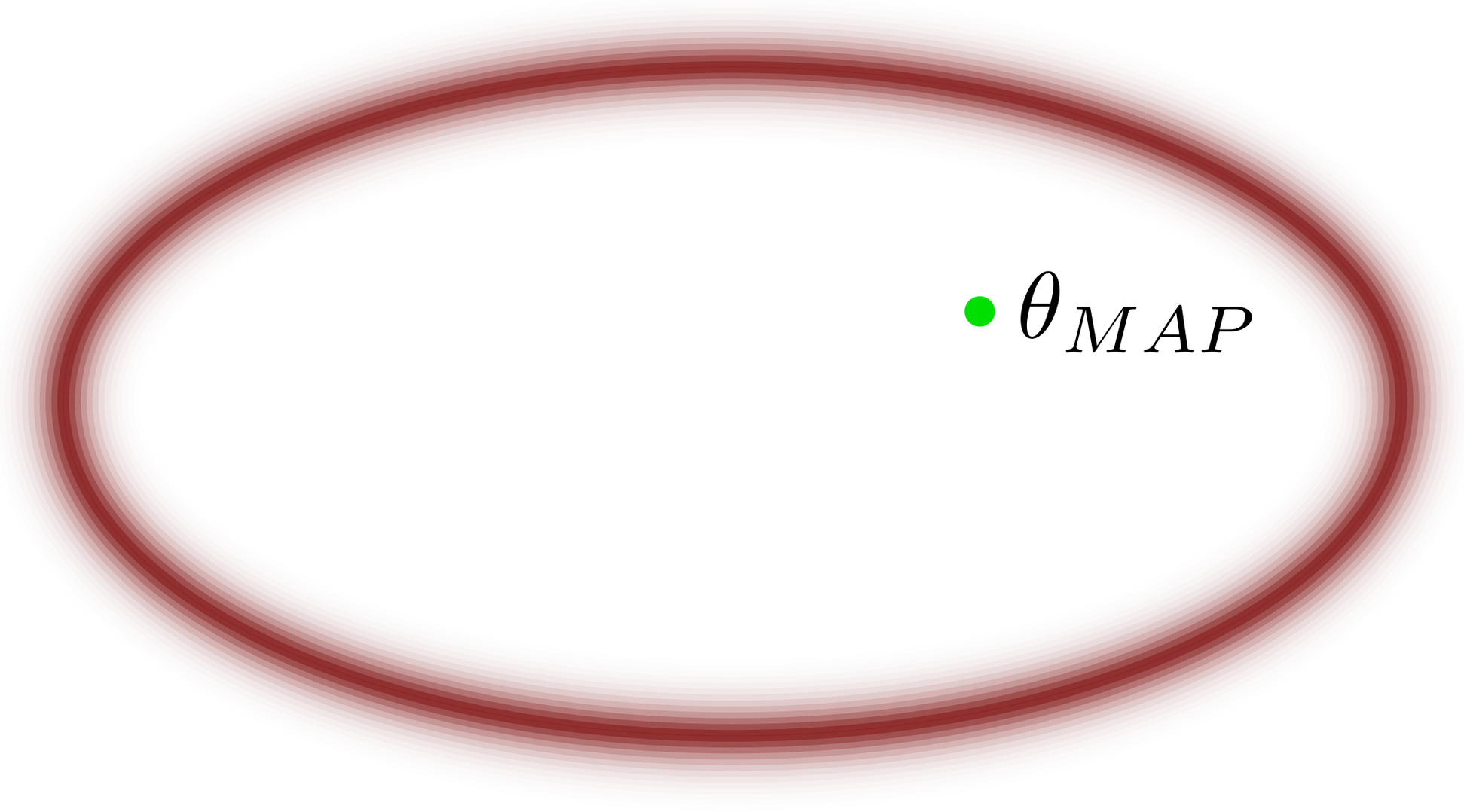
Given extreme symmetries MAP estimators of *some* functions in *some* parameterizations can be accurate.



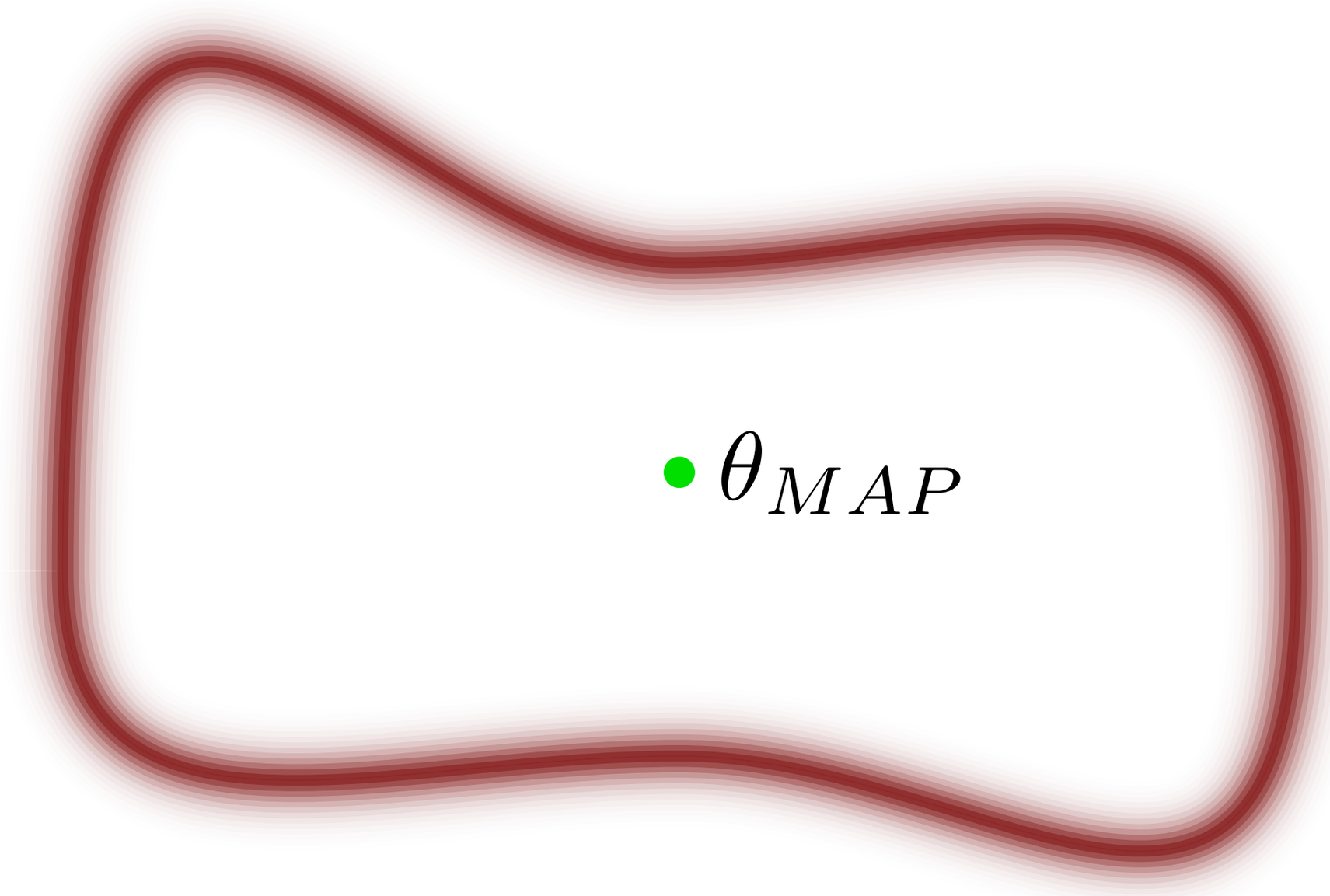
Reparameterizations changes the density and hence the mode, but expectations should be invariant!

$$\pi_S(\phi(\theta)\,|\widetilde{\mathcal{D}}) = \pi_S(\theta|\widetilde{\mathcal{D}}) \left|\frac{\partial \phi}{\partial \theta}\right|$$

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.

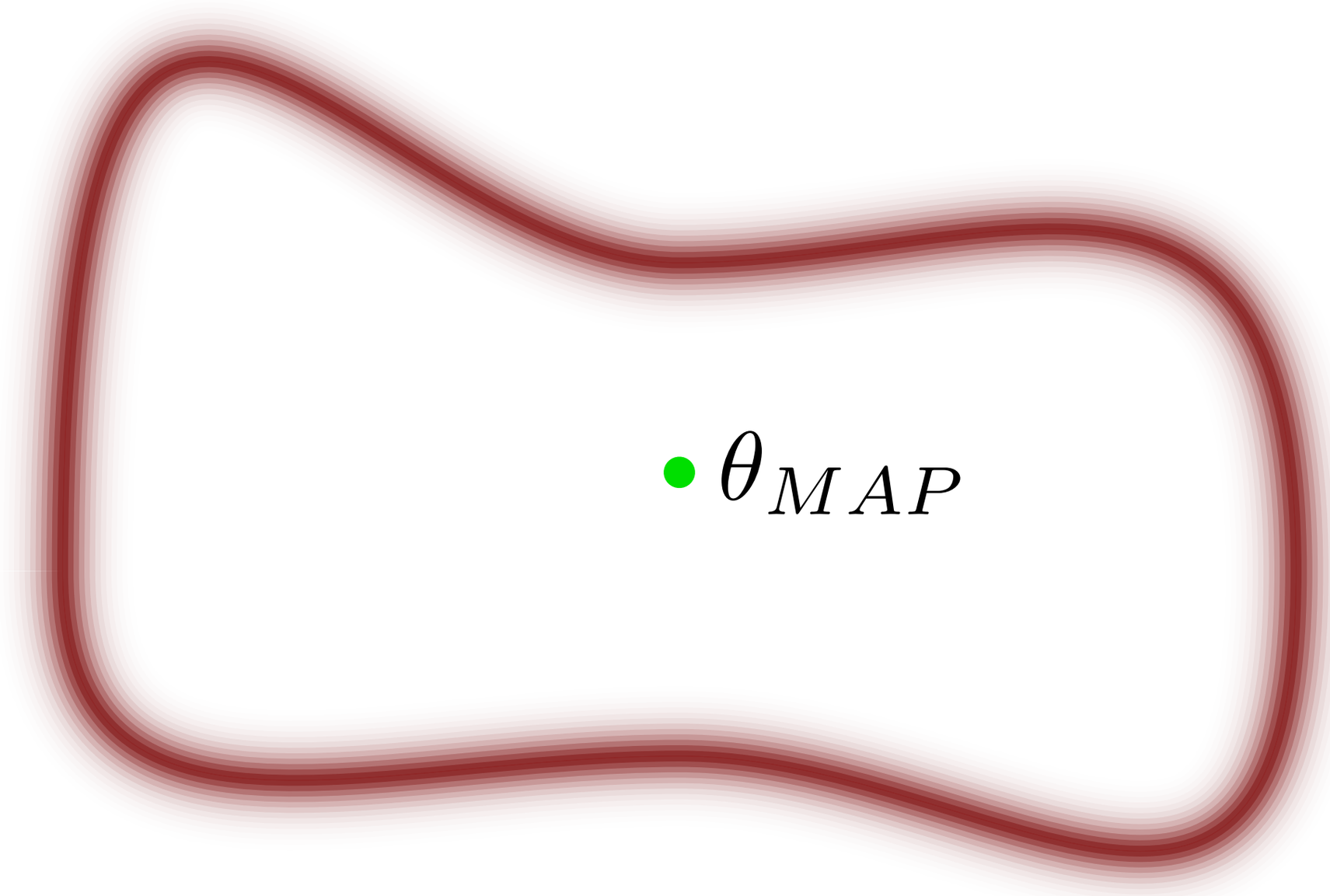


Laplace approximations complement modal estimators with a Hessian to quantify the breadth of the typical set.

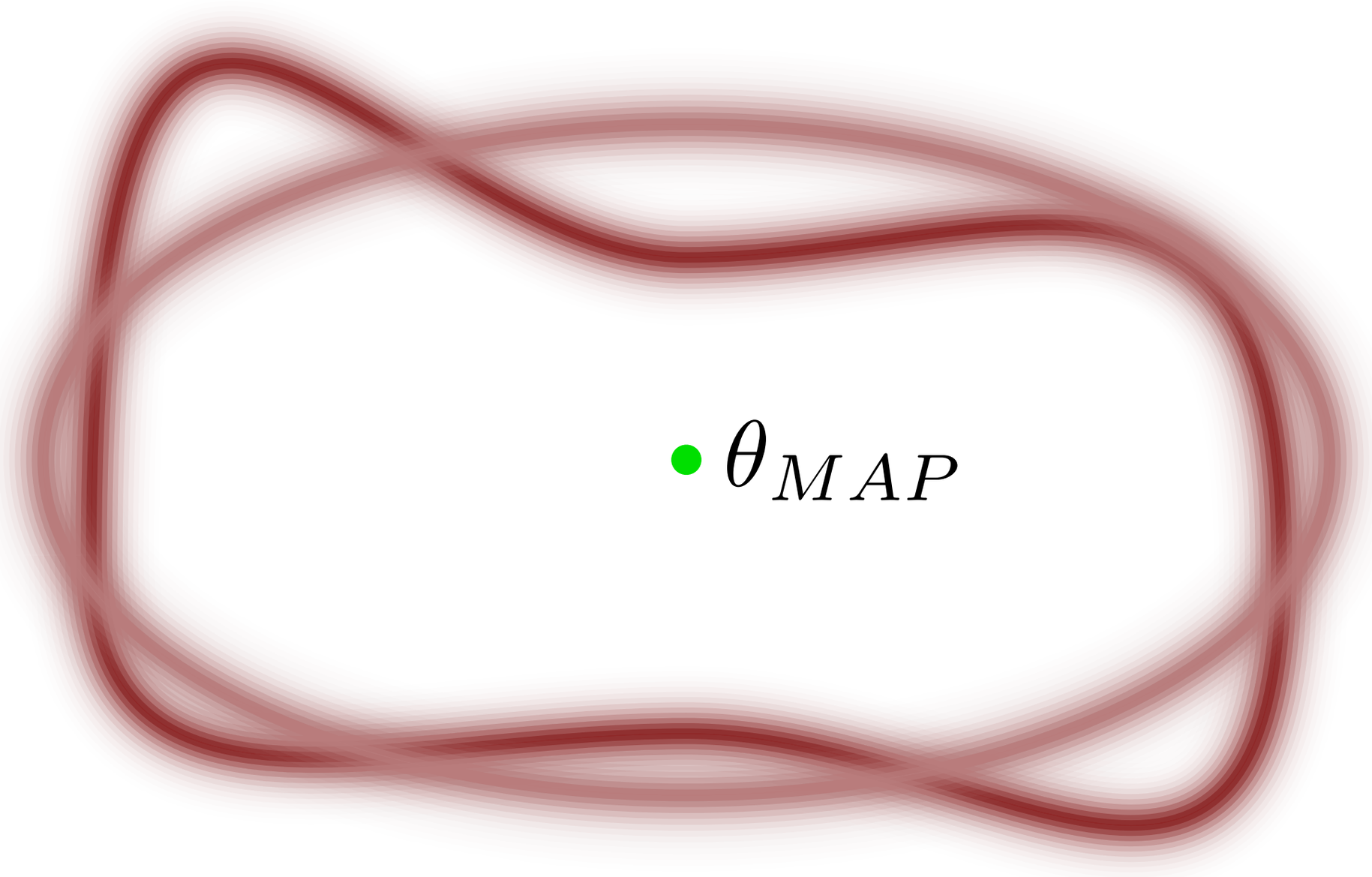
$$\pi_S(\theta|\widetilde{\mathcal{D}}) \approx \mathcal{N}(\theta|\theta_{MAP}, H^{-1})$$

$$H = \left.\frac{\partial^2 \pi_S(\theta|\widetilde{\mathcal{D}})}{\partial \theta^2}\right|_{\theta=\theta_{MAP}}$$

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



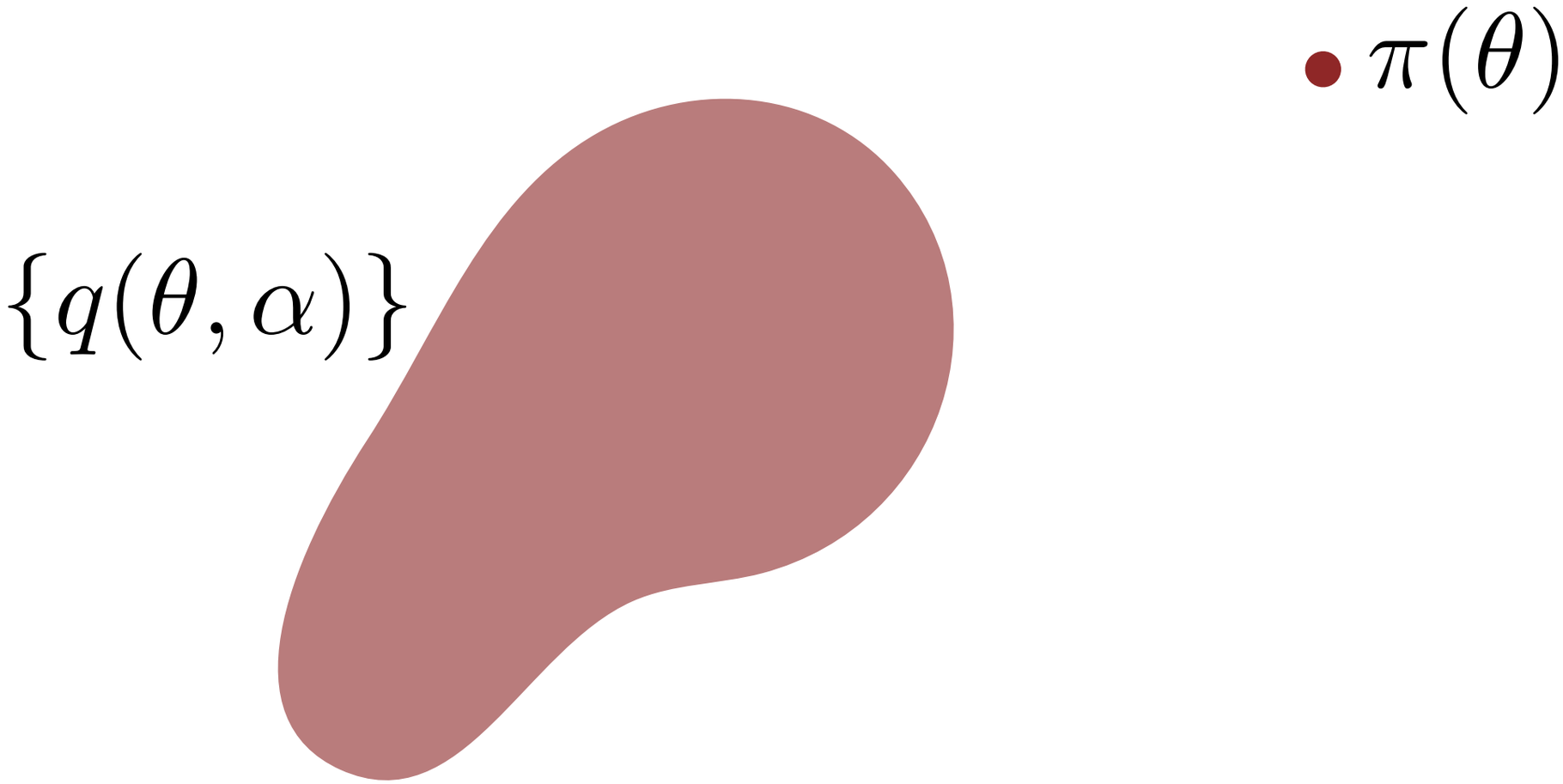
Laplace approximations complement modal estimators with a Hessian to quantify the breadth of the typical set.



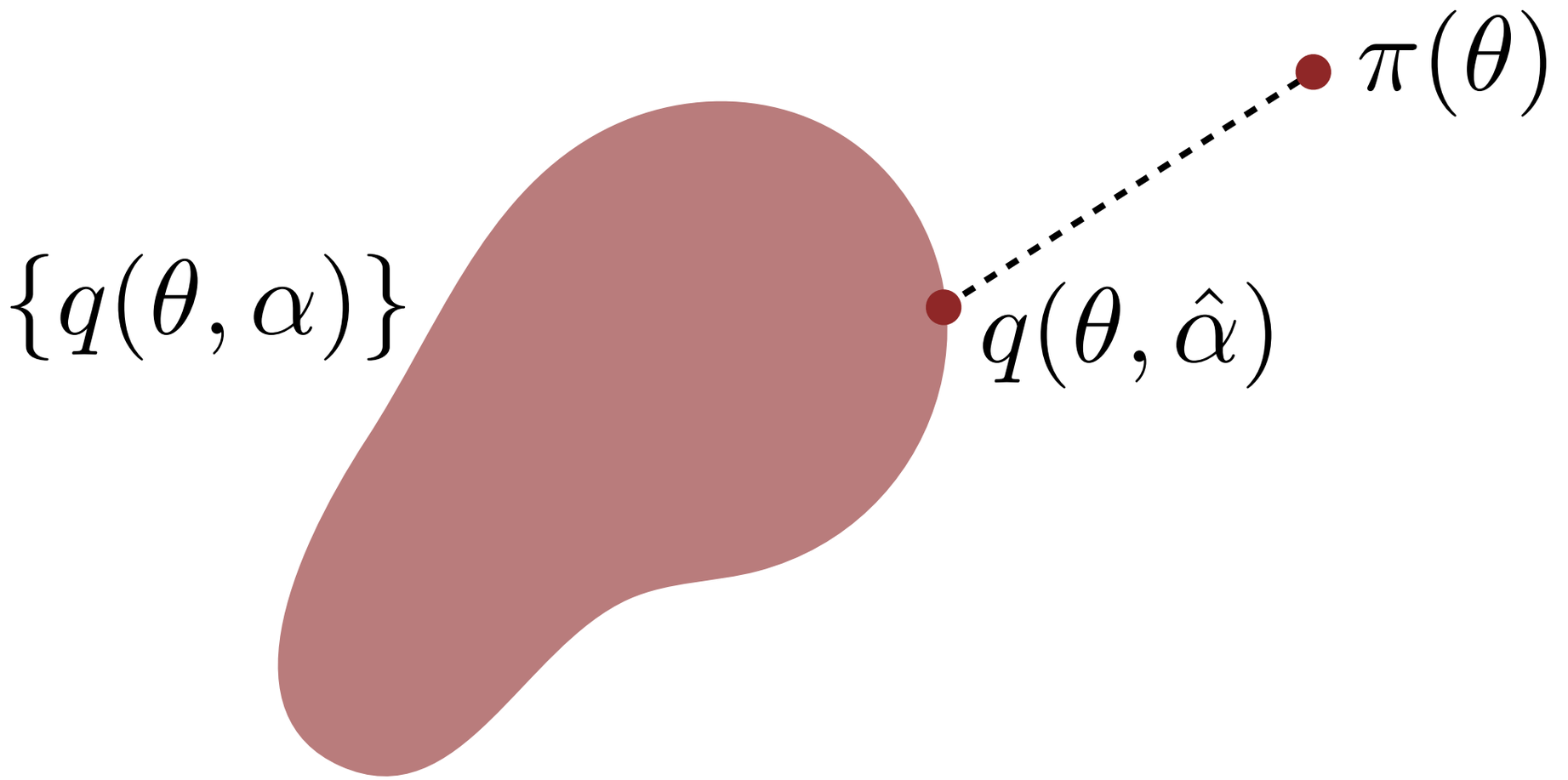
Variational methods turn the approximation problem into a variational optimization.

- $\pi(\theta)$

Variational methods turn the approximation problem into a variational optimization.



Variational methods turn the approximation problem into a variational optimization.



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

$$\pi_S(\theta|\widetilde{\mathcal{D}}) \approx q(\theta, \hat{\alpha})$$

$$\int \mathrm{d}\theta \, \pi_S(\theta|\widetilde{\mathcal{D}}) f(\theta) \approx \int \mathrm{d}\theta \, q(\theta, \hat{\alpha}) f(\theta)$$

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

$$\pi_S(\theta|\widetilde{\mathcal{D}})$$

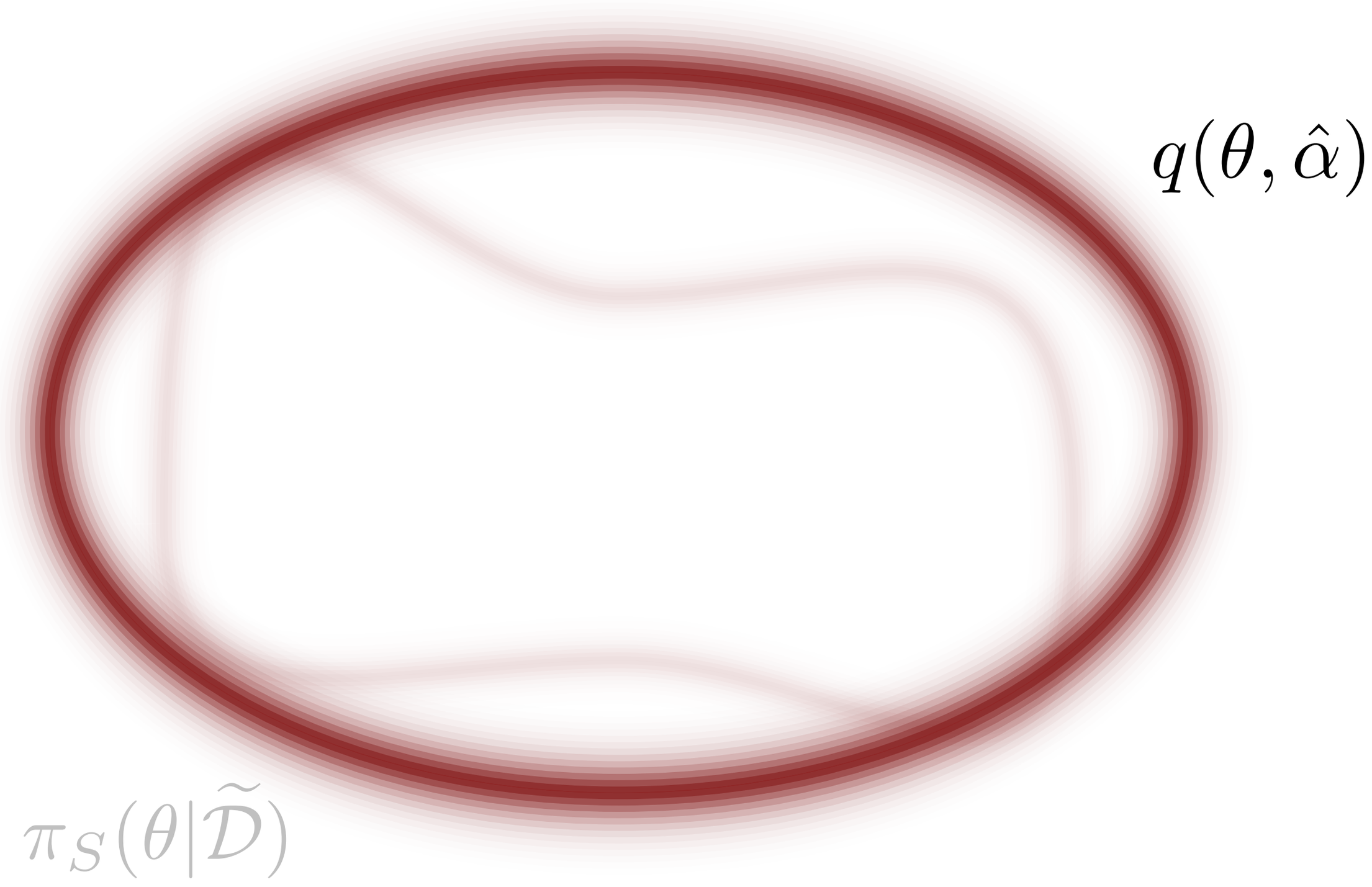
And then use the closet element of the variational family to approximate posterior expectations.

$\pi_S(\theta|\widetilde{\mathcal{D}})$

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

$$\pi_S(\theta|\widetilde{\mathcal{D}})$$

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



Stochastic methods construct random estimators of the exact expectations.

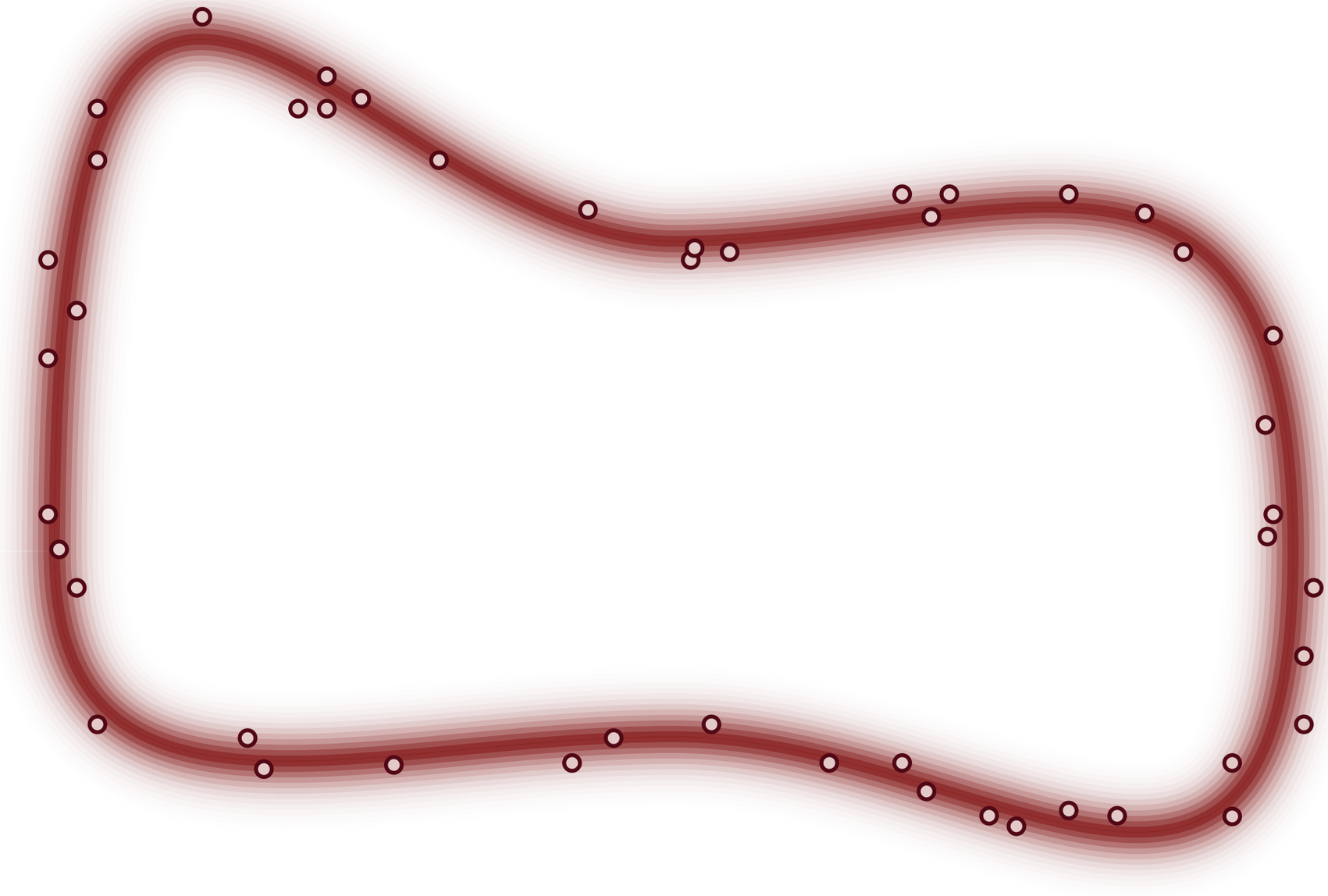
$$\{\theta_1, \ldots, \theta_N\}$$

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

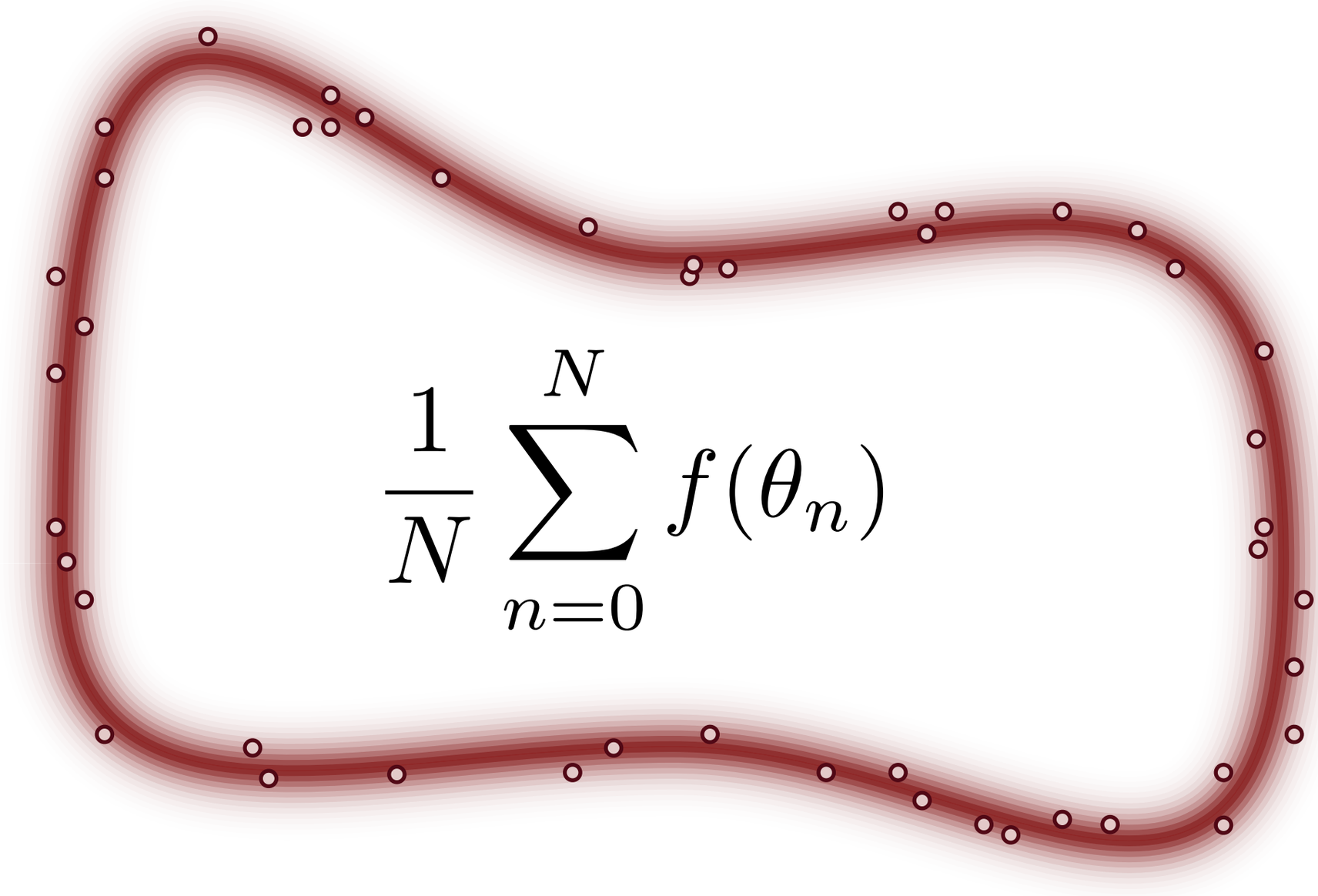
The foremost stochastic techniques are *Monte Carlo* methods.

Here integrals are approximated with Monte Carlo estimators using samples drawn from the posterior.

Here integrals are approximated with Monte Carlo estimators using samples drawn from the posterior.



These Monte Carlo estimators are consistent and eventually converge to the true expectation.



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

These Monte Carlo estimators are consistent and eventually converge to the true expectation.

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

For a scalable algorithm, these samples have to be generated with a Markov chain.

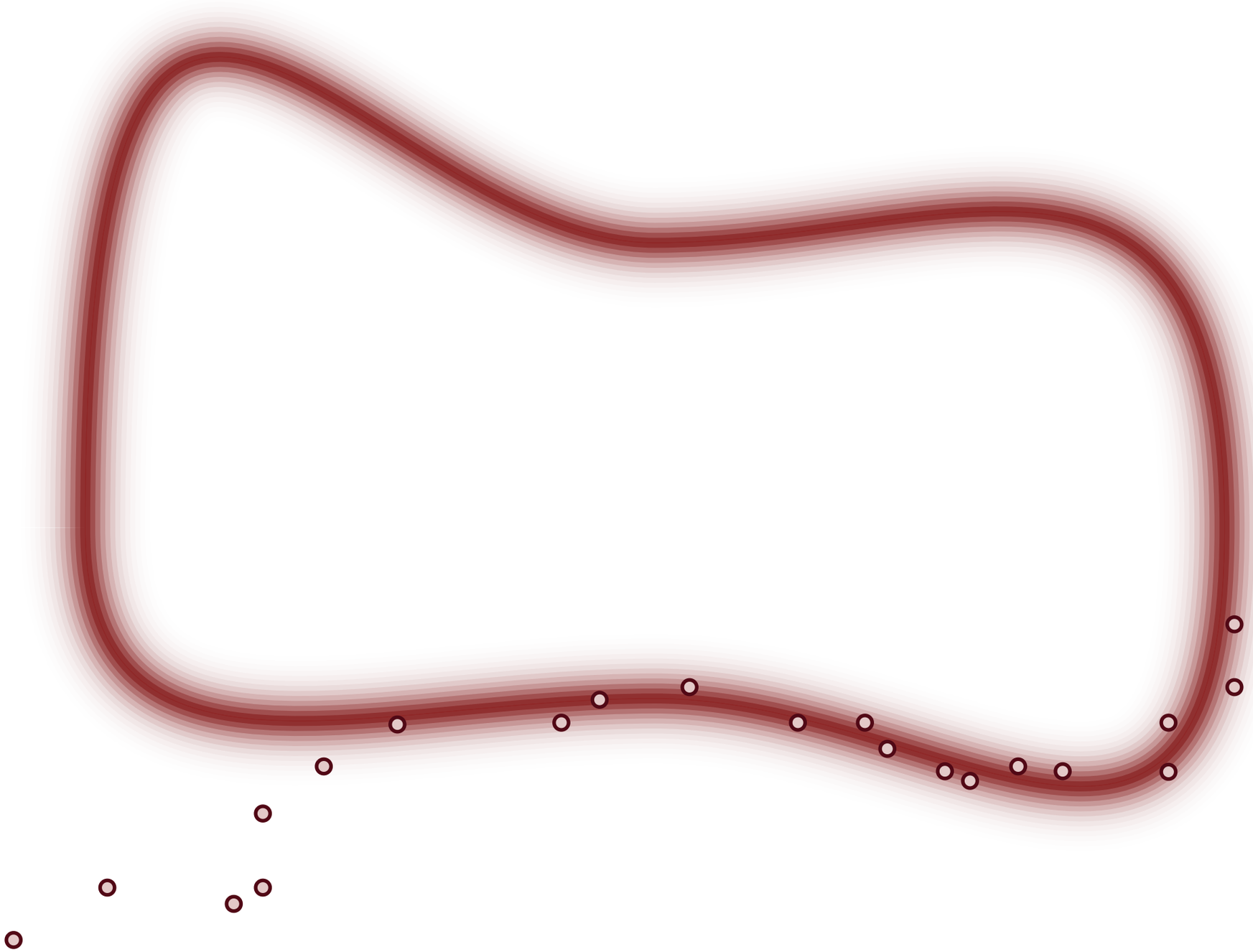
$$T(\theta, \theta')$$

For a scalable algorithm, these samples have to be generated with a Markov chain.

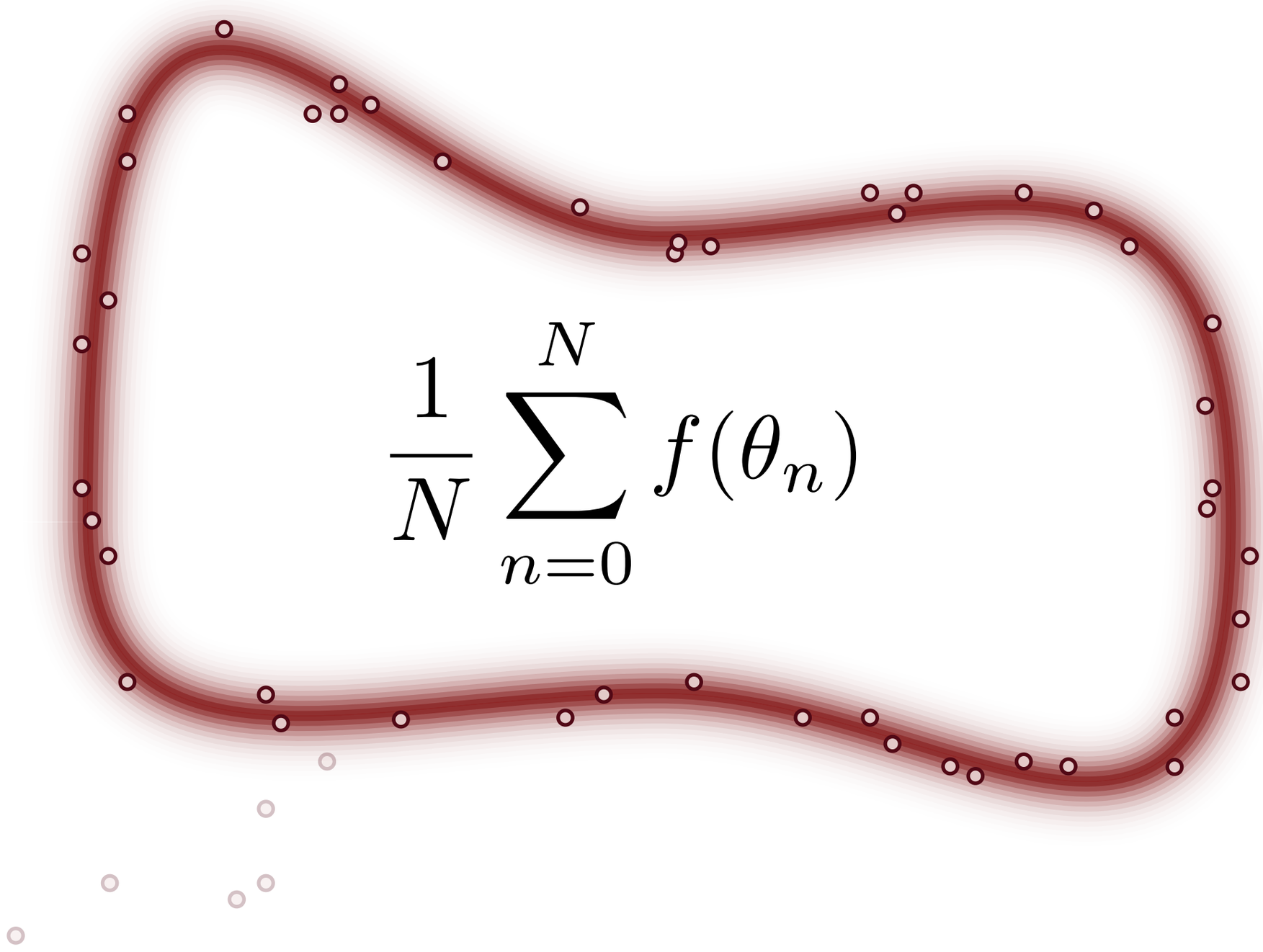
$$\pi_S(\theta|\widetilde{\mathcal{D}}) = \int \mathrm{d}\theta'\, T(\theta,\theta')\pi_S(\theta'|\widetilde{\mathcal{D}})$$

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*.

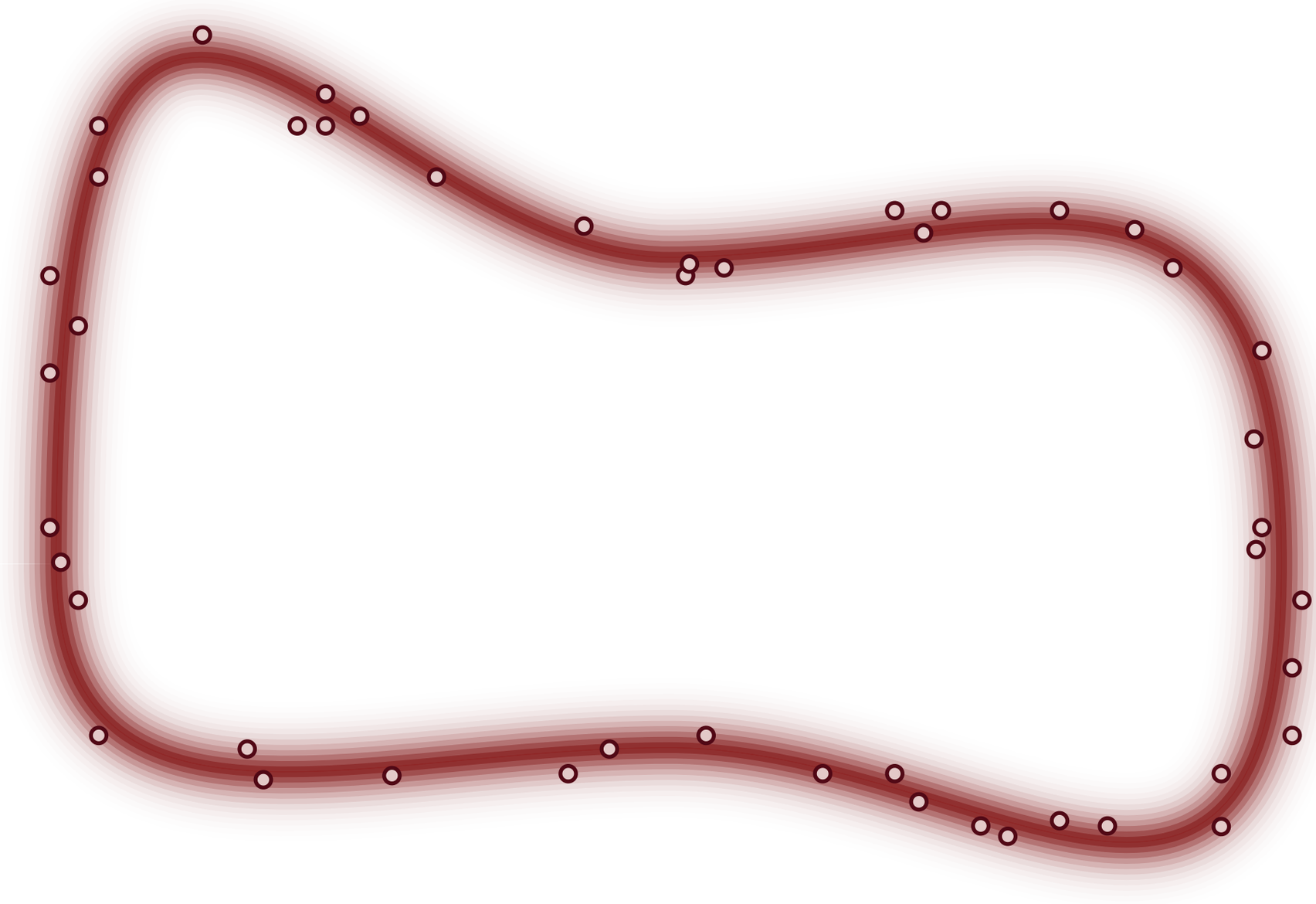


$$\frac{1}{N} \sum_{n=0}^{N} f(\theta_n)$$

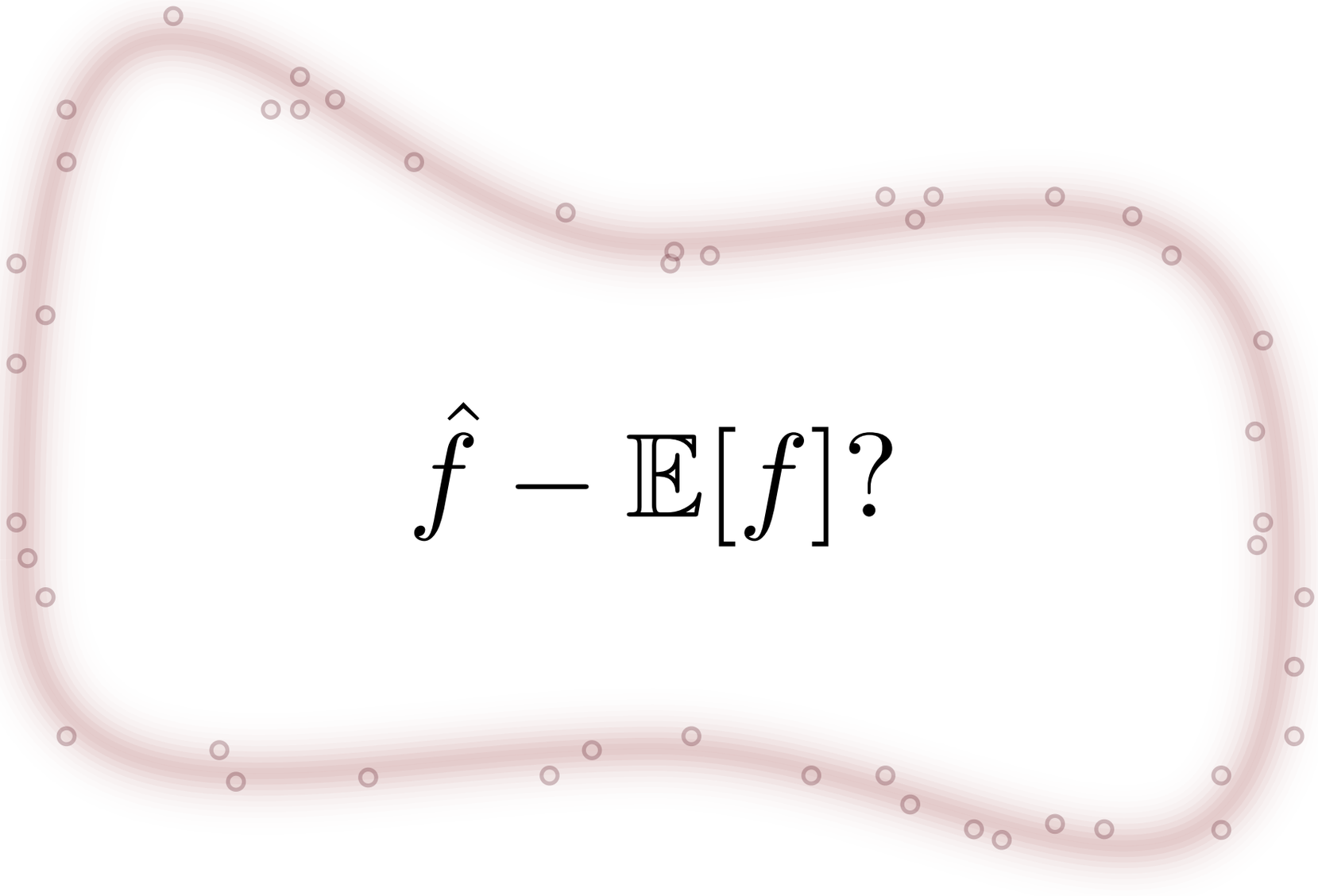
If run long enough the Markov chain defines consistent *Markov Chain Monte Carlo estimators.*

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

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

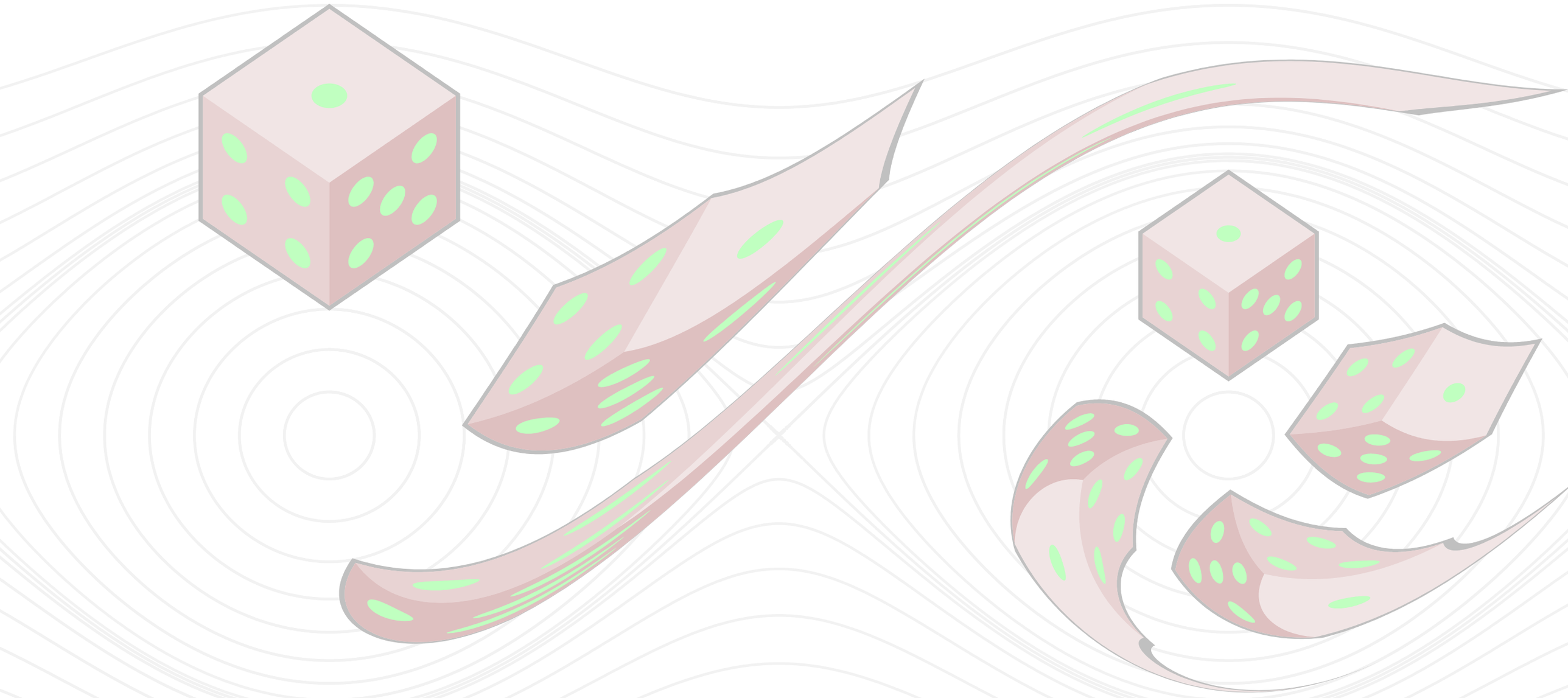


To use any method responsibly, however, you have to be able to quantify the accuracy of the estimation!



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

Backups



Marginalization

Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.

$$\pi_S(\theta_1, \theta_2, \ldots, \theta_n | \widetilde{\mathcal{D}})$$

Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.

$$\pi_S(\theta_1, \theta_2, \ldots, \theta_n | \widetilde{\mathcal{D}})$$

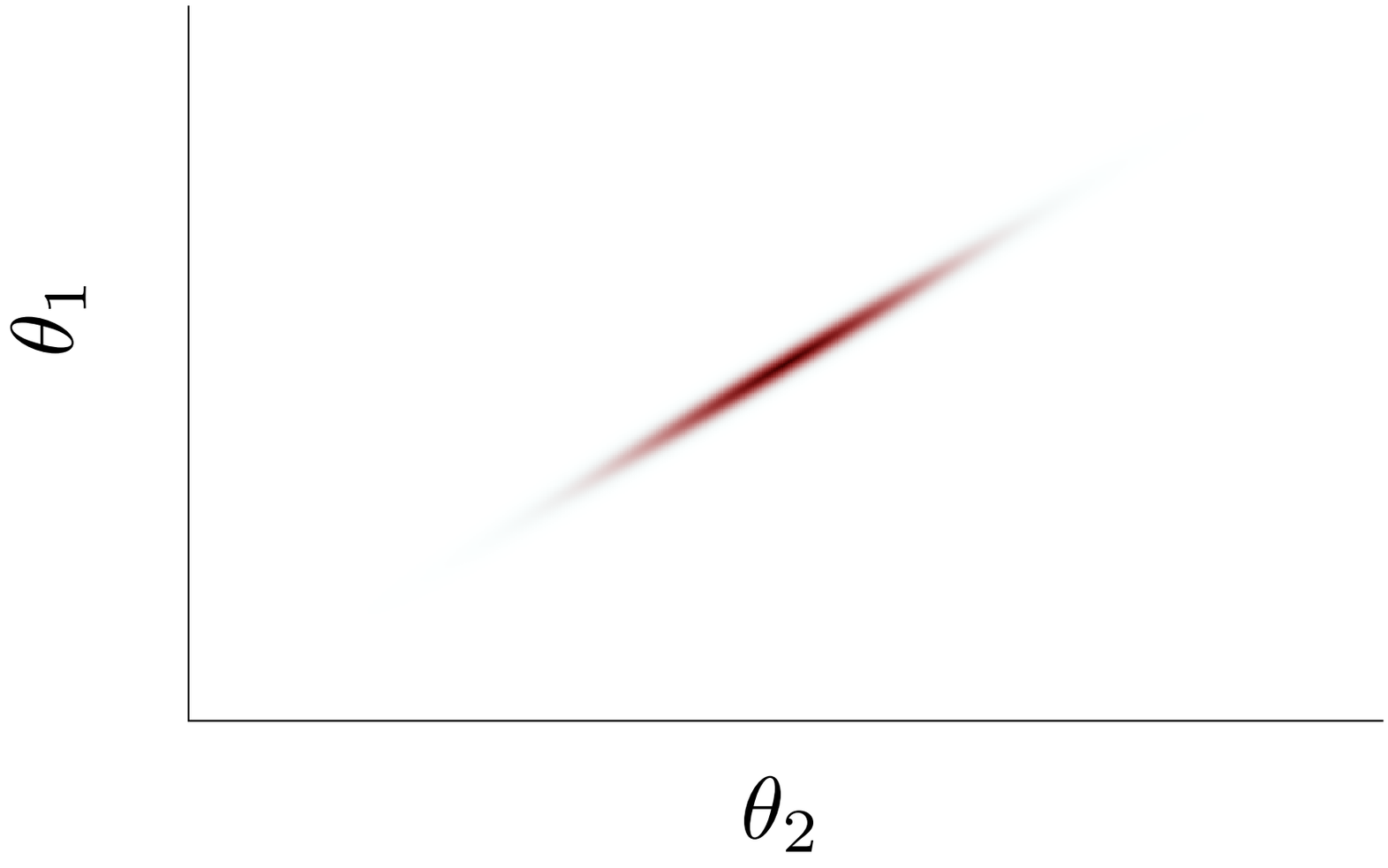
Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.

$$\pi_S(\theta_1, \theta_2, \ldots, \theta_n | \widetilde{\mathcal{D}})$$

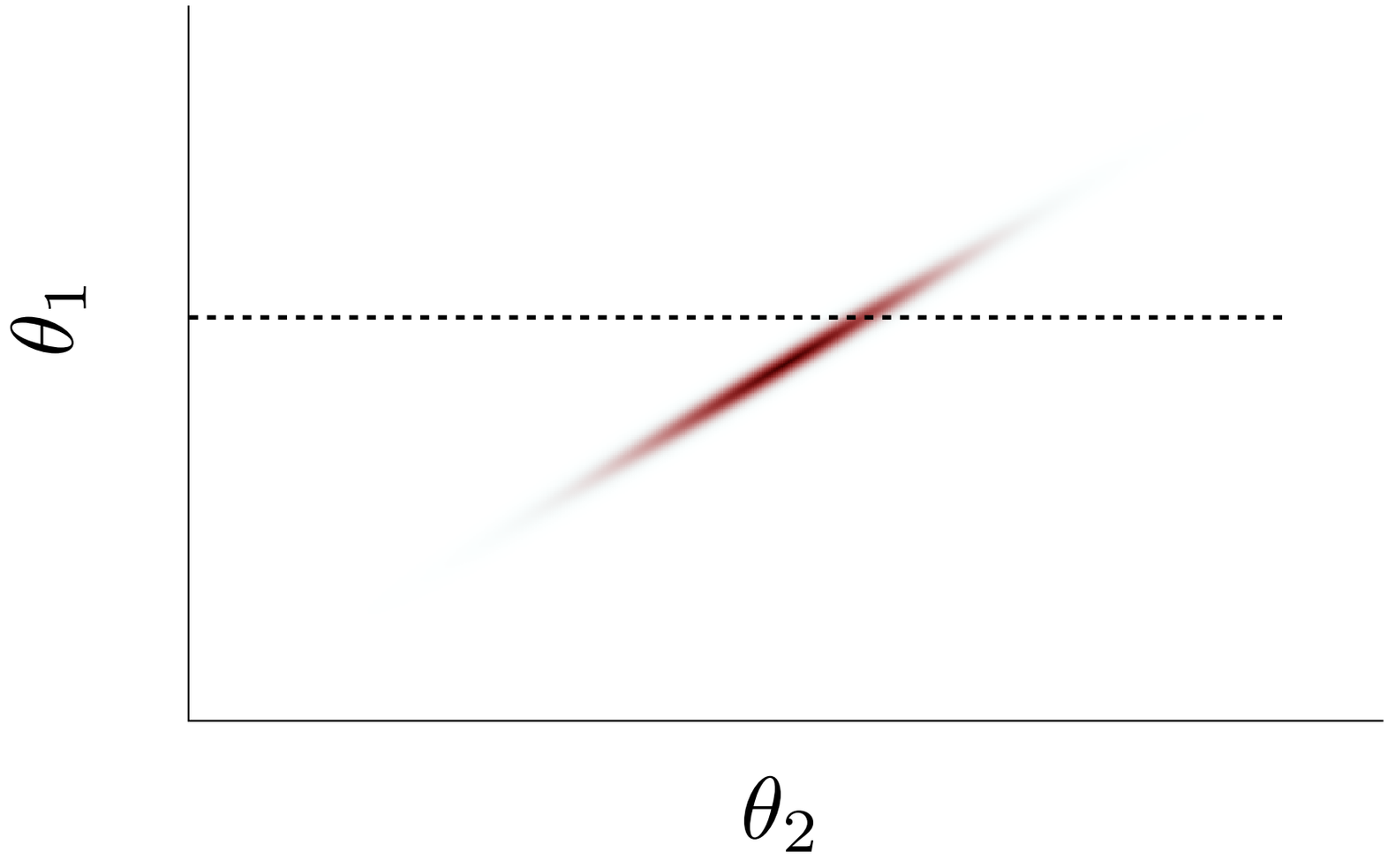
Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.

$$\pi_S(\theta_2, \ldots, \theta_n|\widetilde{\mathcal{D}}) = \int \mathrm{d}\theta_1 \, \pi_S(\theta_1, \theta_2, \ldots, \theta_n|\widetilde{\mathcal{D}})$$

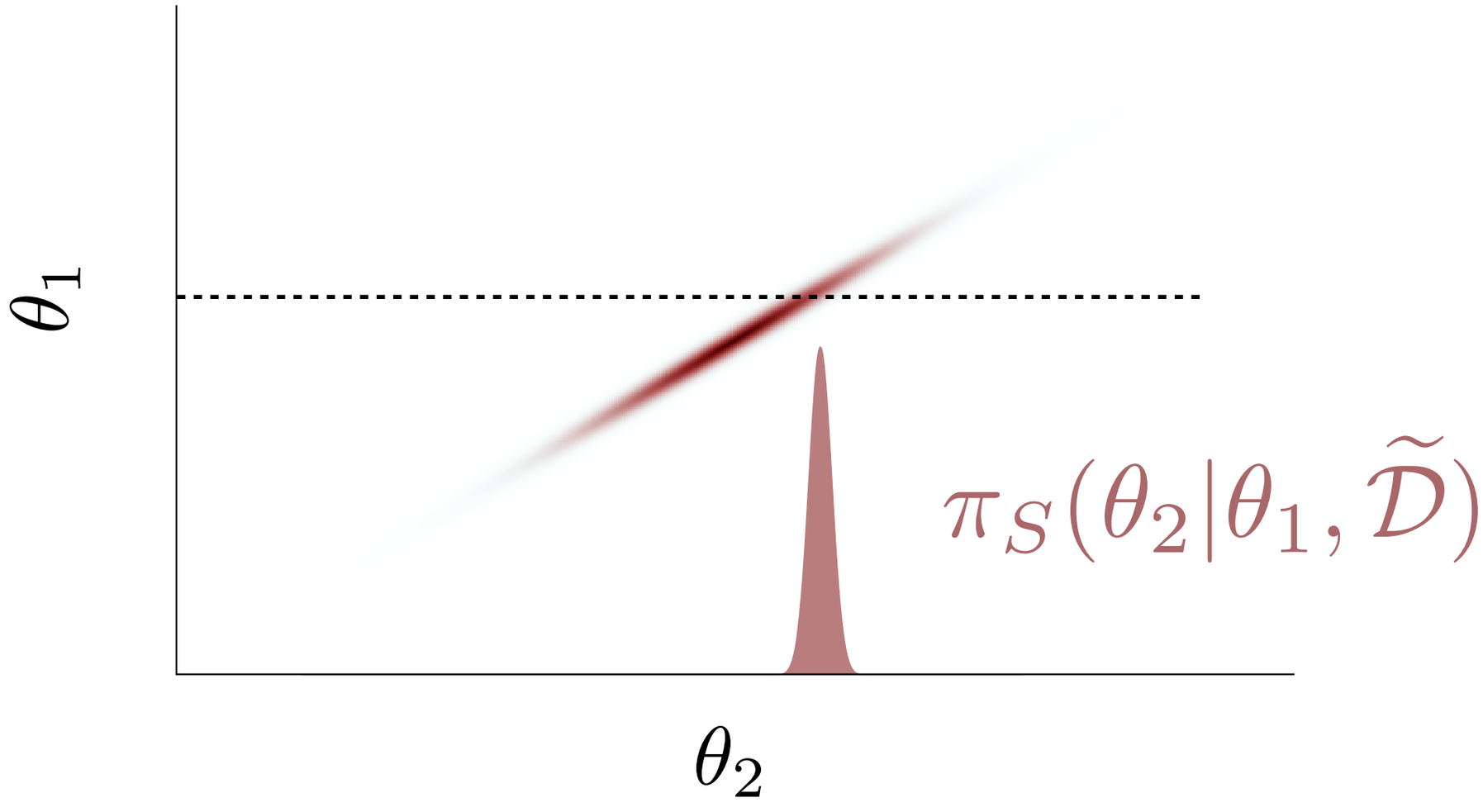
Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.



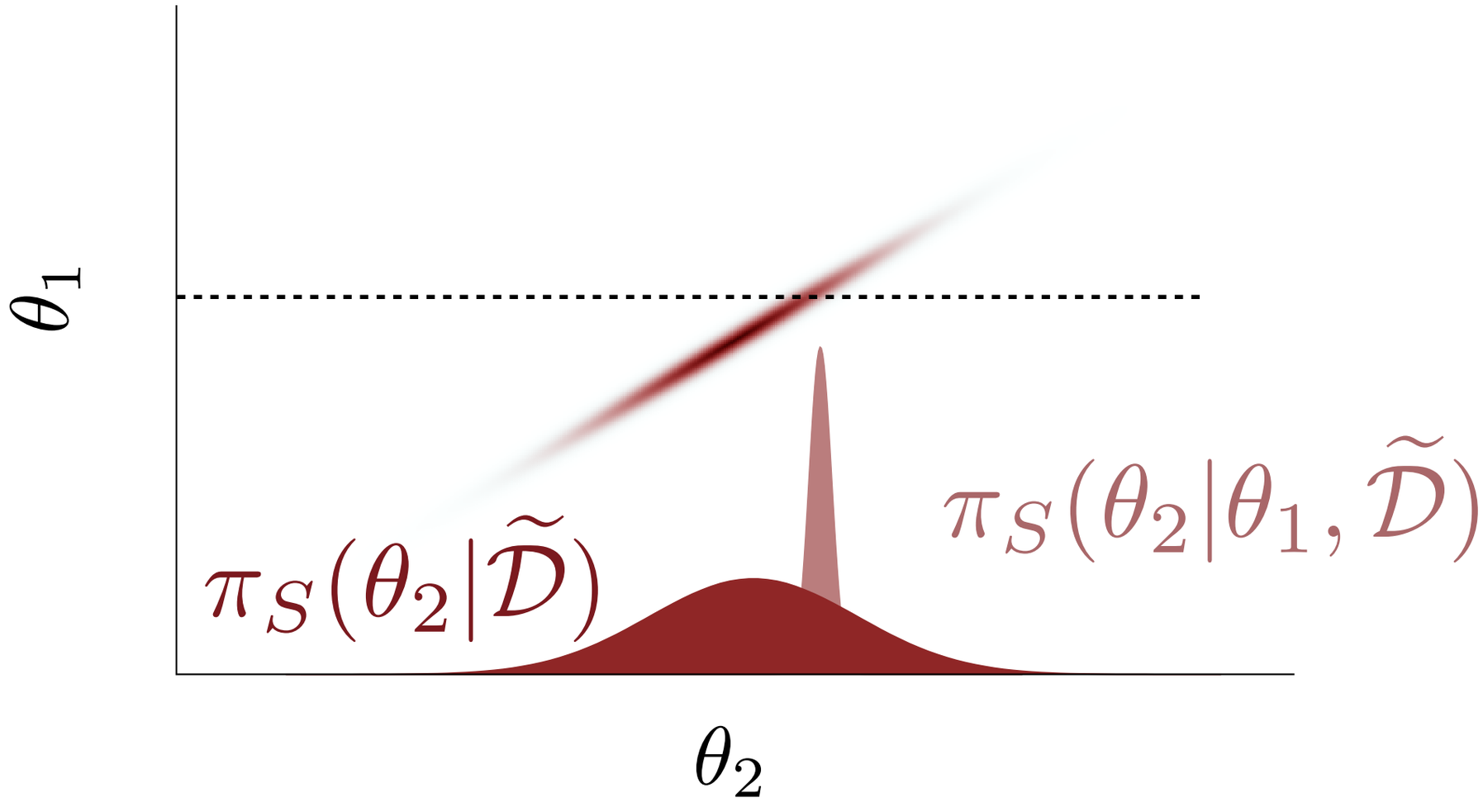
Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.



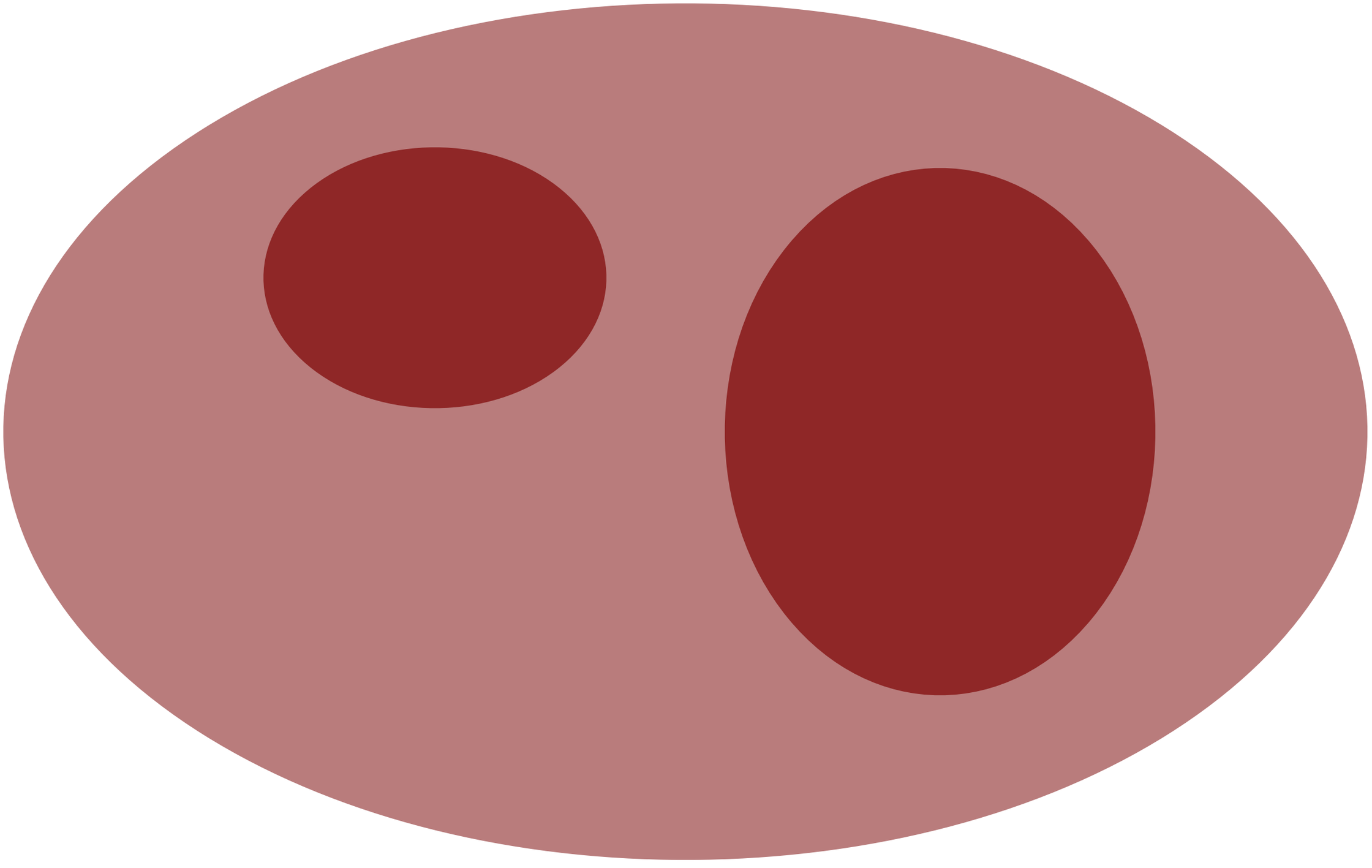
Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.



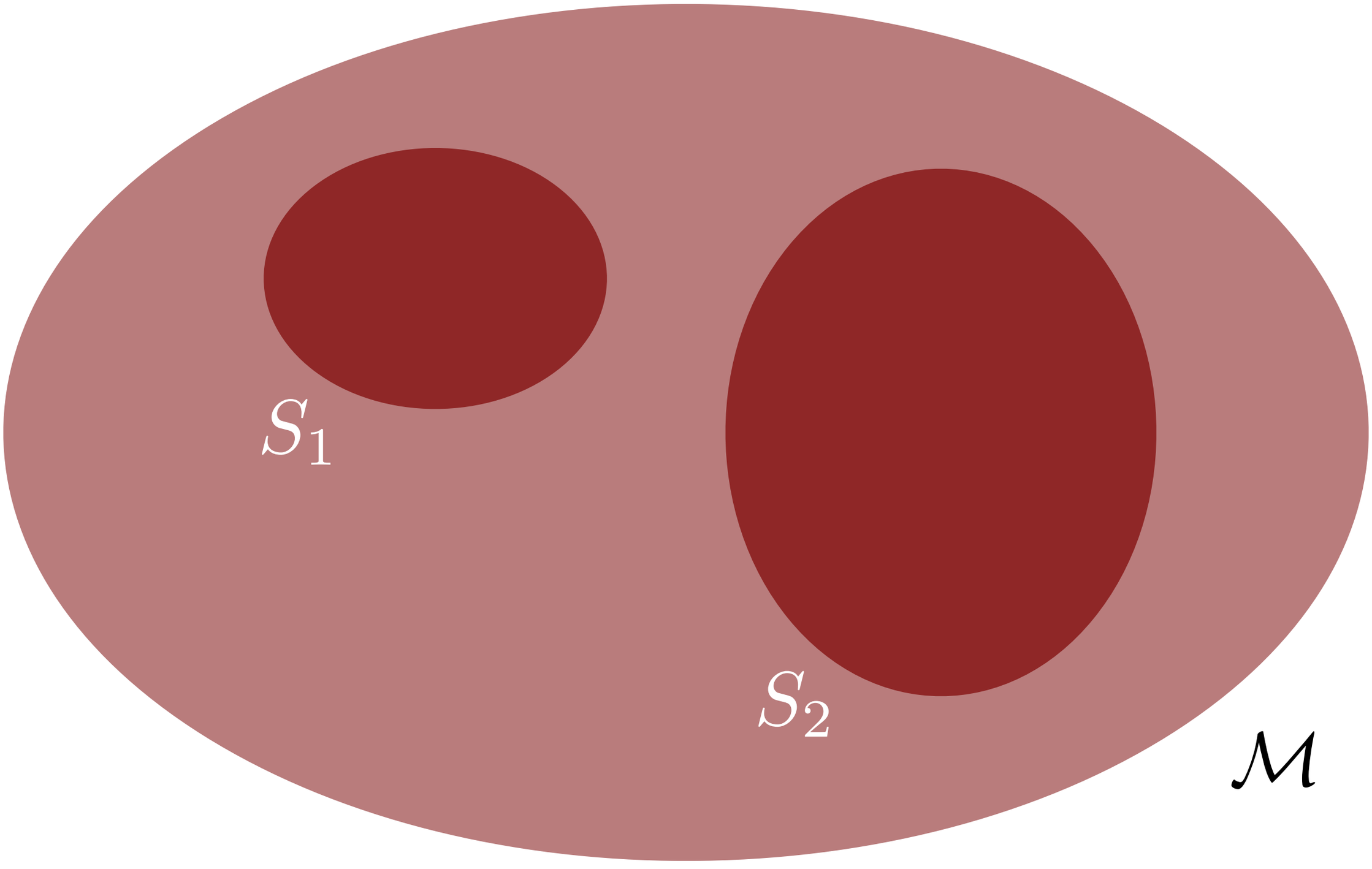
Systematic uncertainties are incorporated by modeling nuisance parameters and marginalizing them out.



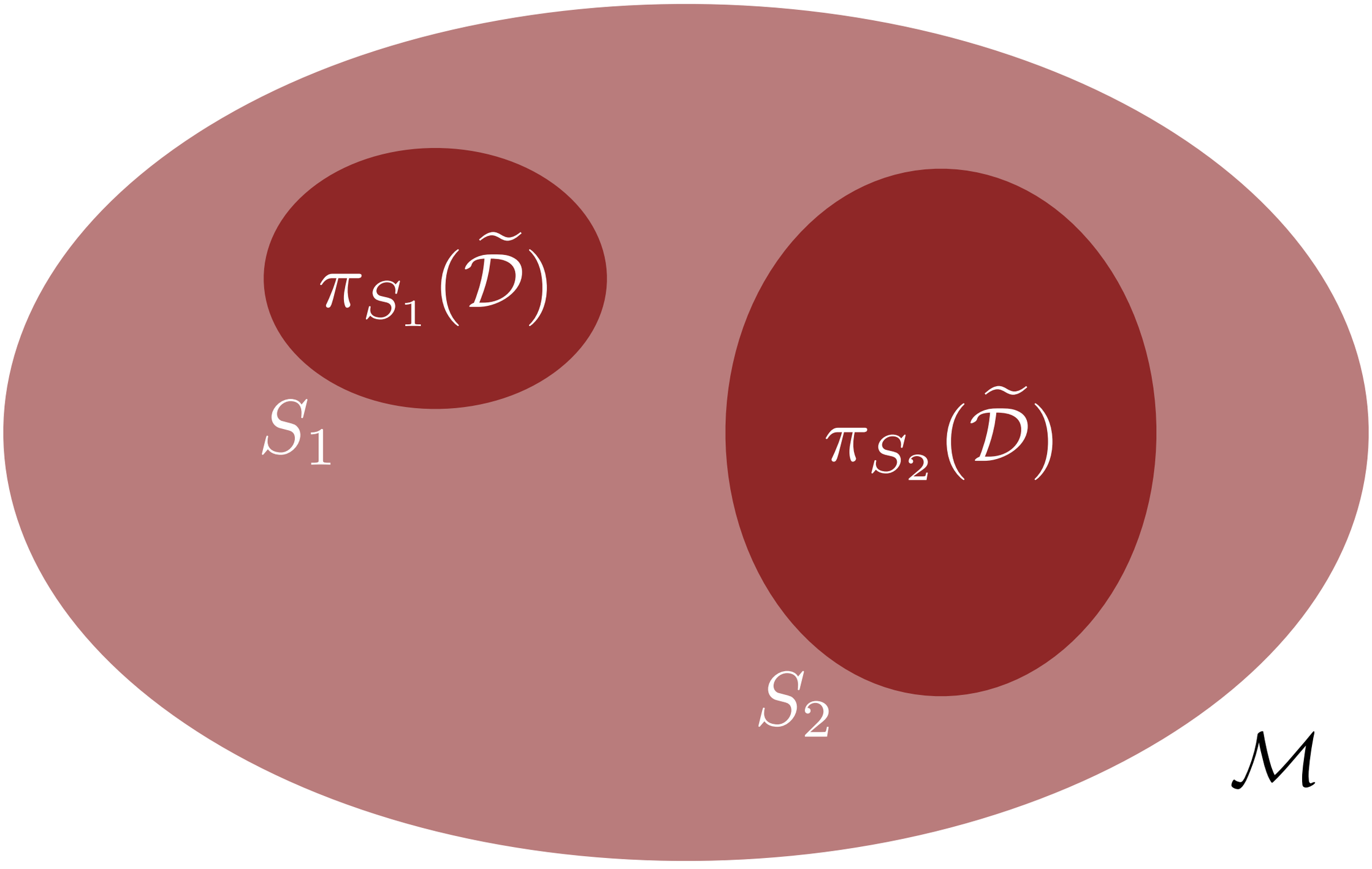
Model Comparison



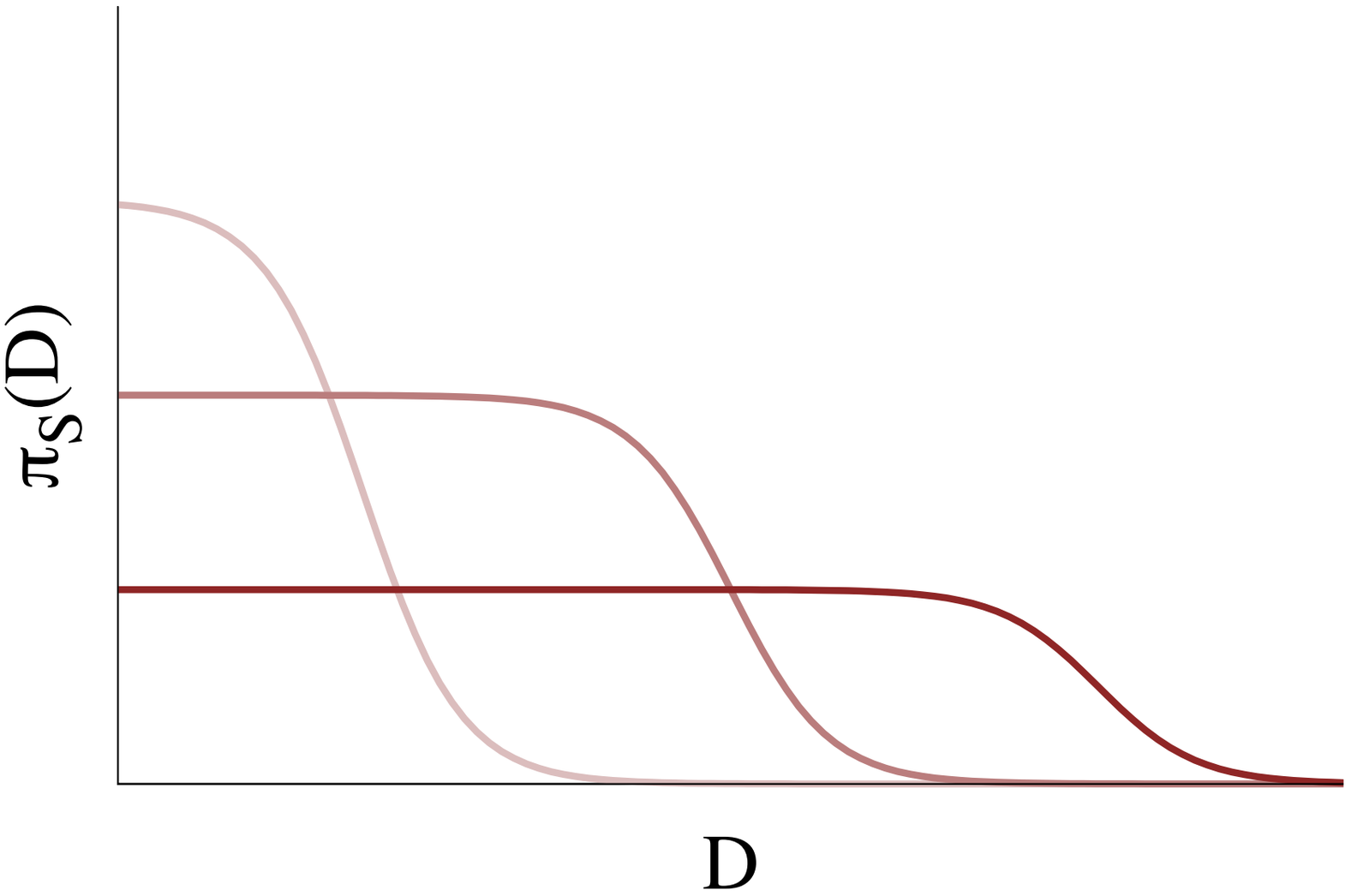
We can also compare different subsets of models.



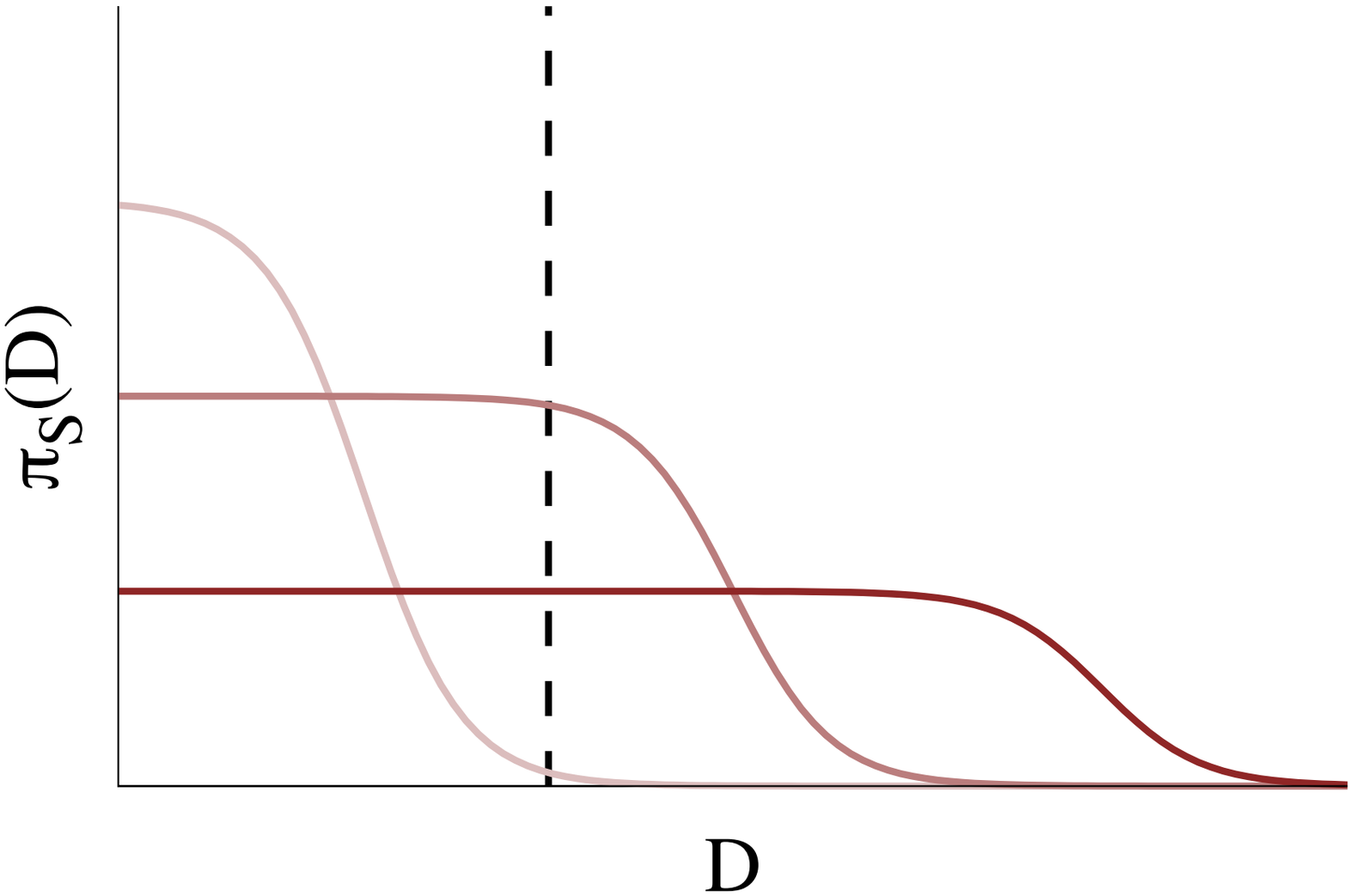
In particular, the marginal likelihood defines a measure of model comparison.



Such comparisons are beneficial because the marginal likelihood naturally incorporates Occam's Razor.



Such comparisons are beneficial because the marginal likelihood naturally incorporates Occam's Razor.

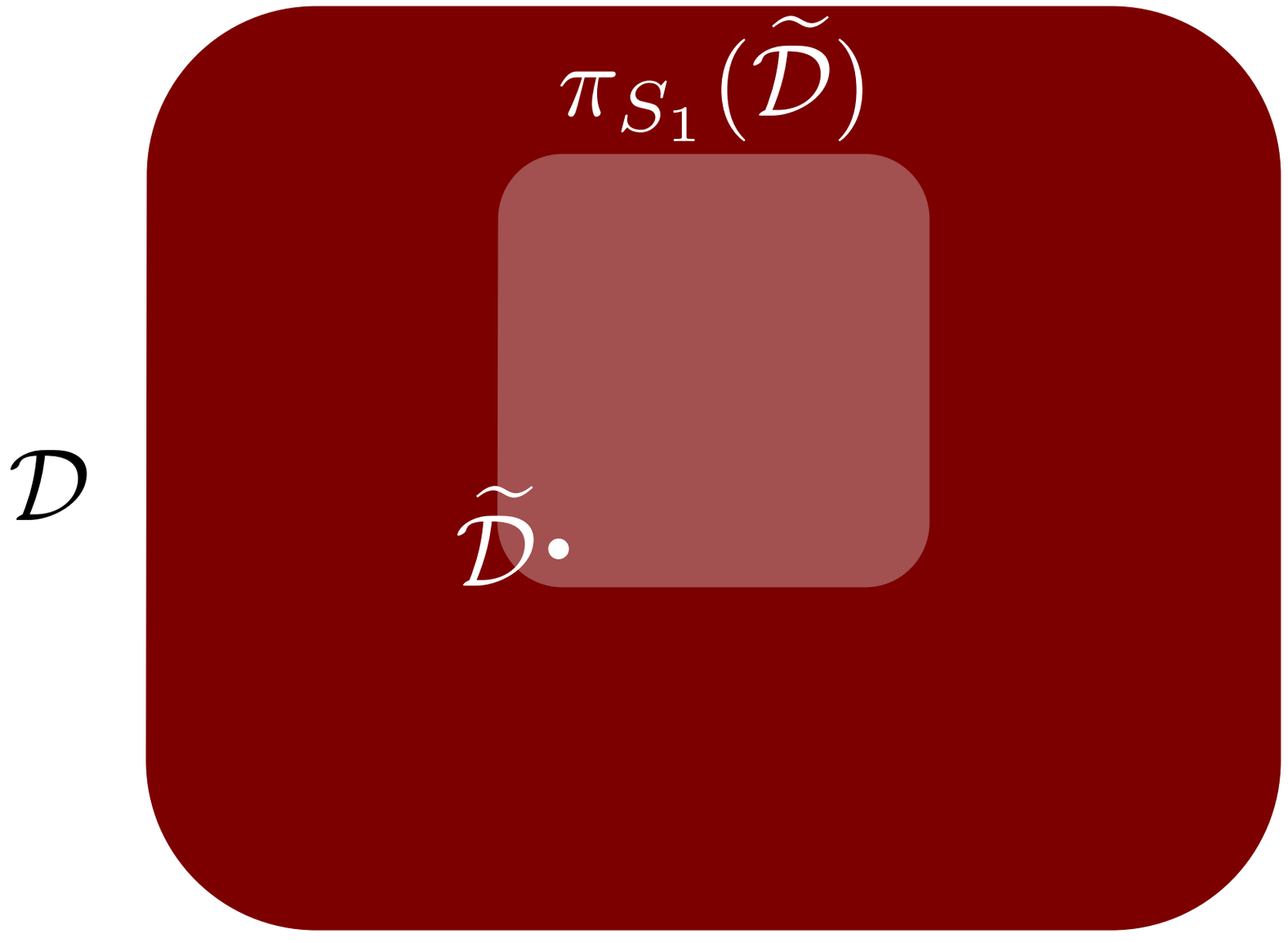


Such comparisons are beneficial because the marginal likelihood naturally incorporates Occam's Razor.

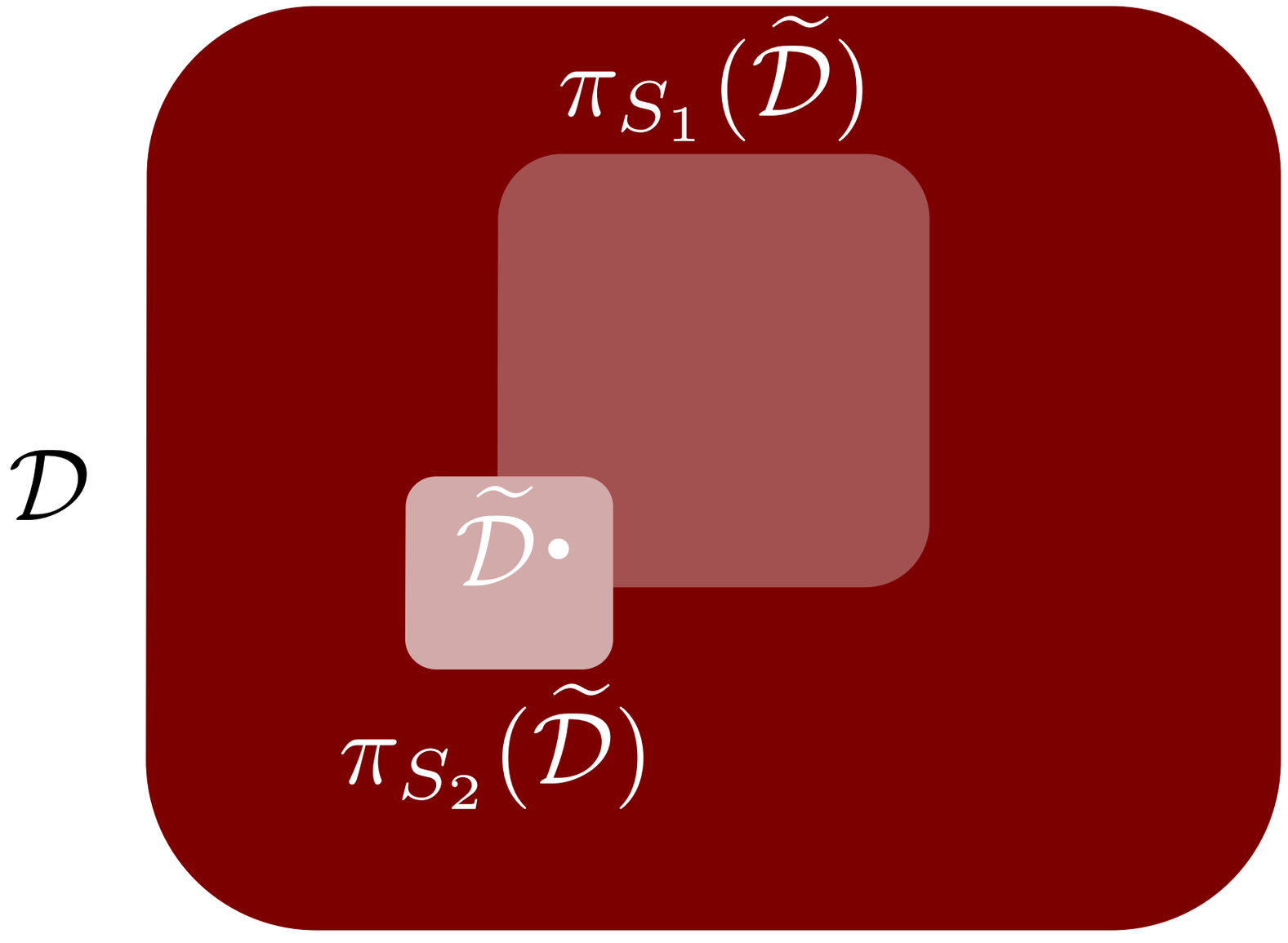
$\mathcal{D}$

$\widetilde{\mathcal{D}}$.

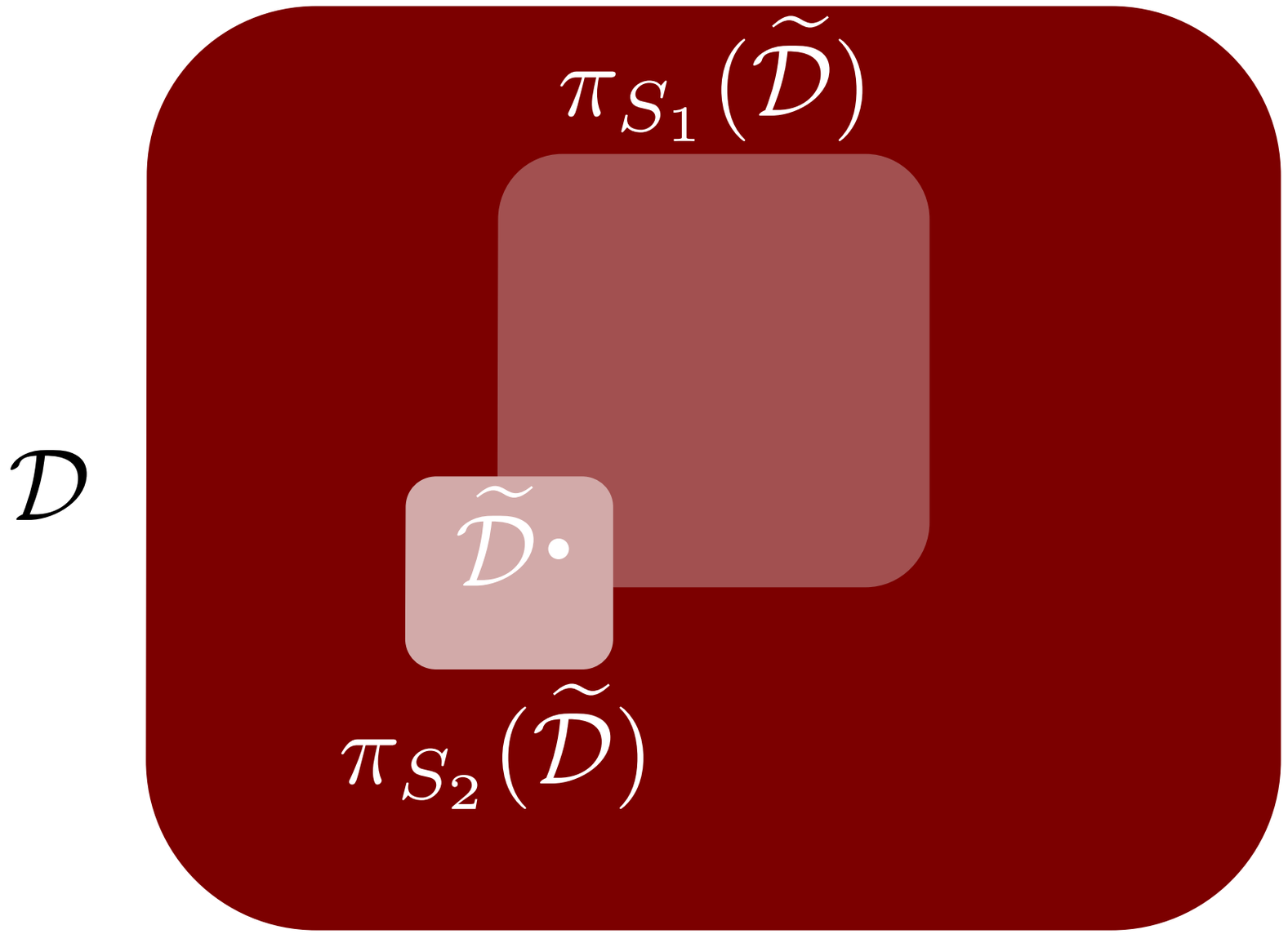
Such comparisons are beneficial because the marginal likelihood naturally incorporates Occam's Razor.



Such comparisons are beneficial because the marginal likelihood naturally incorporates Occam's Razor.



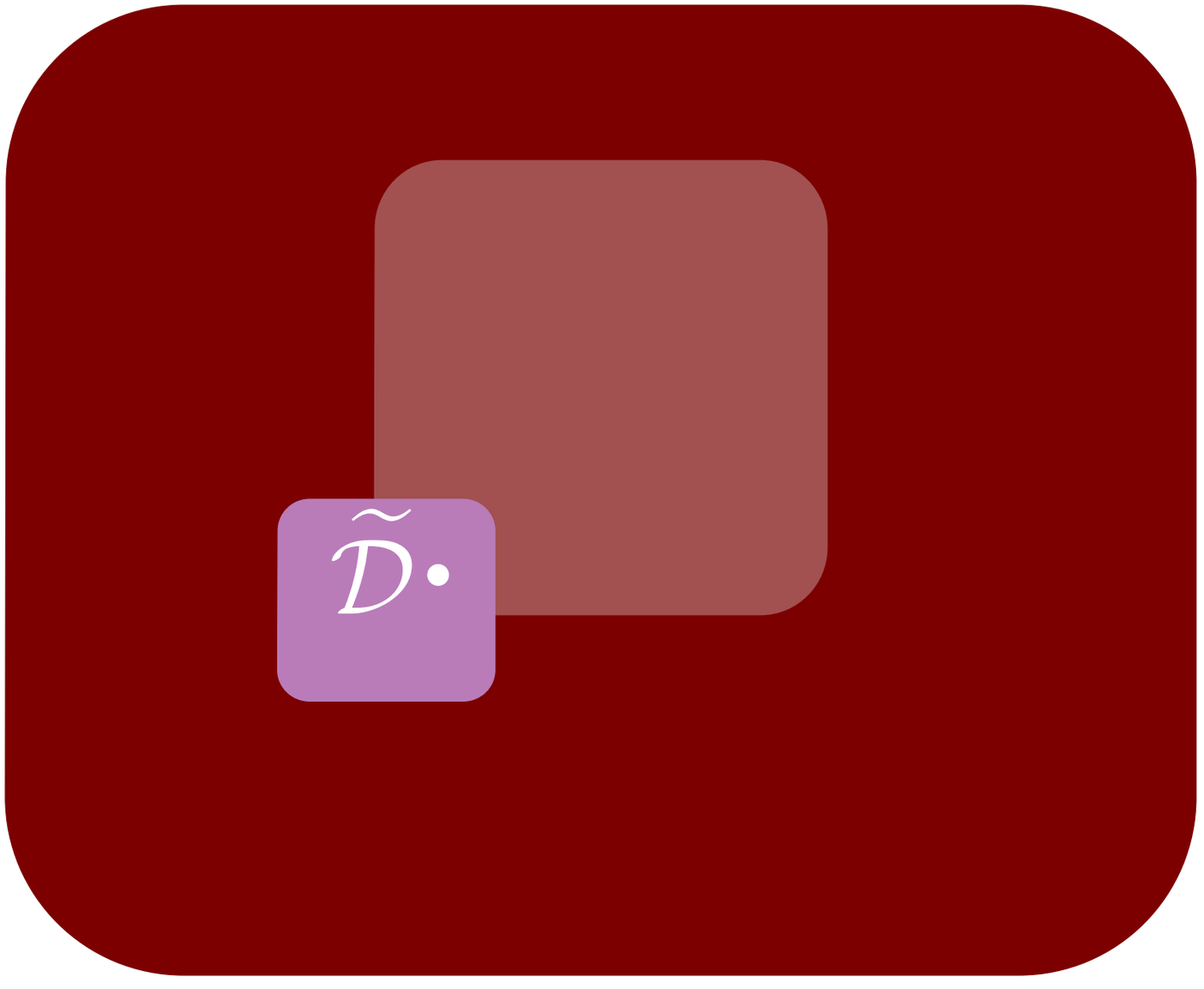
Such comparisons are beneficial because the marginal likelihood naturally incorporates Occam's Razor.



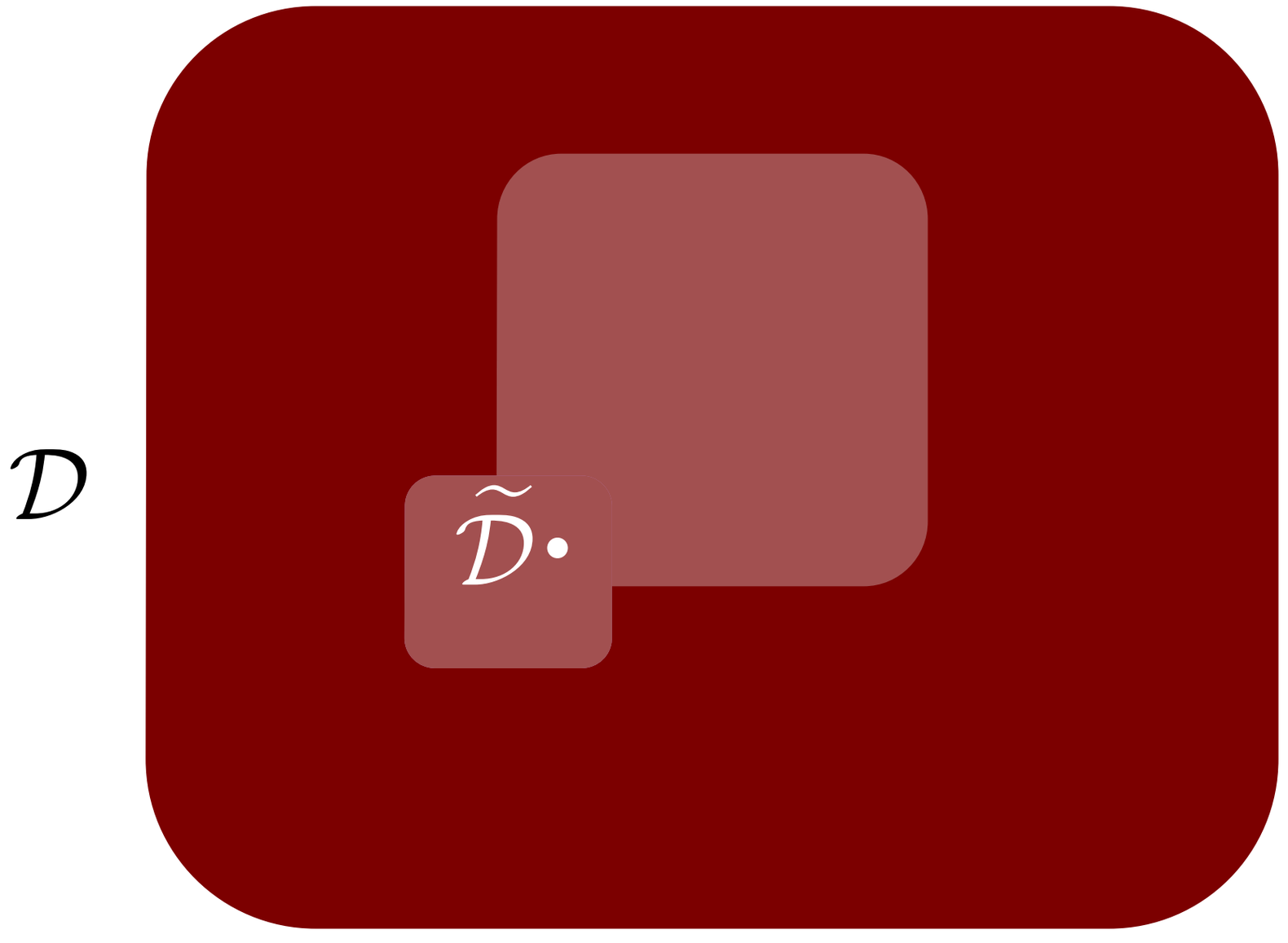
$$\pi_{S_2}(\widetilde{\mathcal{D}}) > \pi_{S_1}(\widetilde{\mathcal{D}})$$

Embracing uncertainty, we can always average over the spaces instead of selecting just one.

$\mathcal{D}$



Embracing uncertainty, we can always average over the spaces instead of selecting just one.

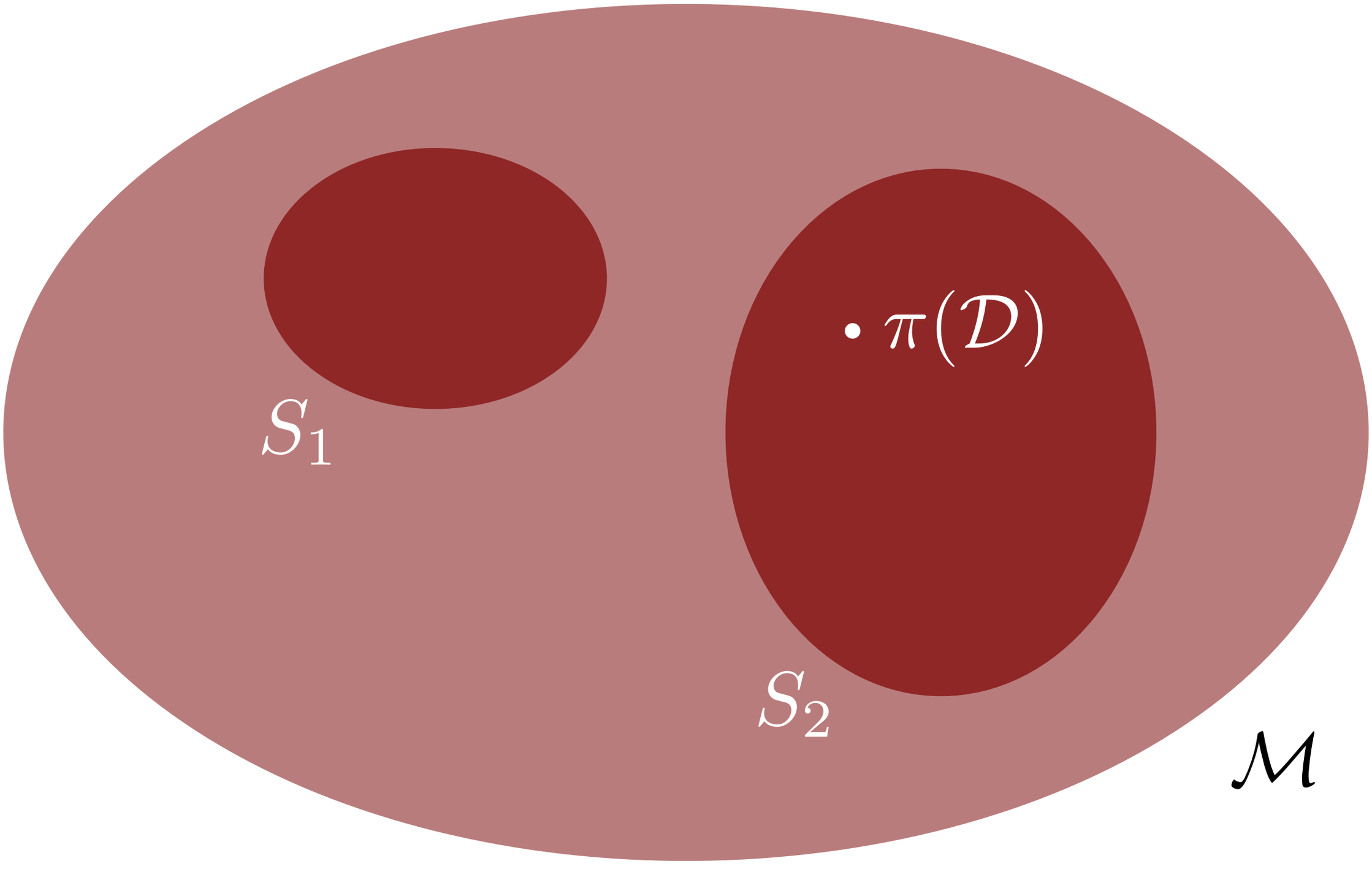


$$\pi_{S_1+S_2}(\theta|\widetilde{\mathcal{D}}) = \frac{\pi_{S_1}(\widetilde{\mathcal{D}})\pi_{S_1}(\theta|\widetilde{\mathcal{D}}) + \pi_{S_2}(\widetilde{\mathcal{D}})\pi_{S_2}(\theta|\widetilde{\mathcal{D}})}{\pi_{S_1}(\widetilde{\mathcal{D}}) + \pi_{S_2}(\widetilde{\mathcal{D}})}$$

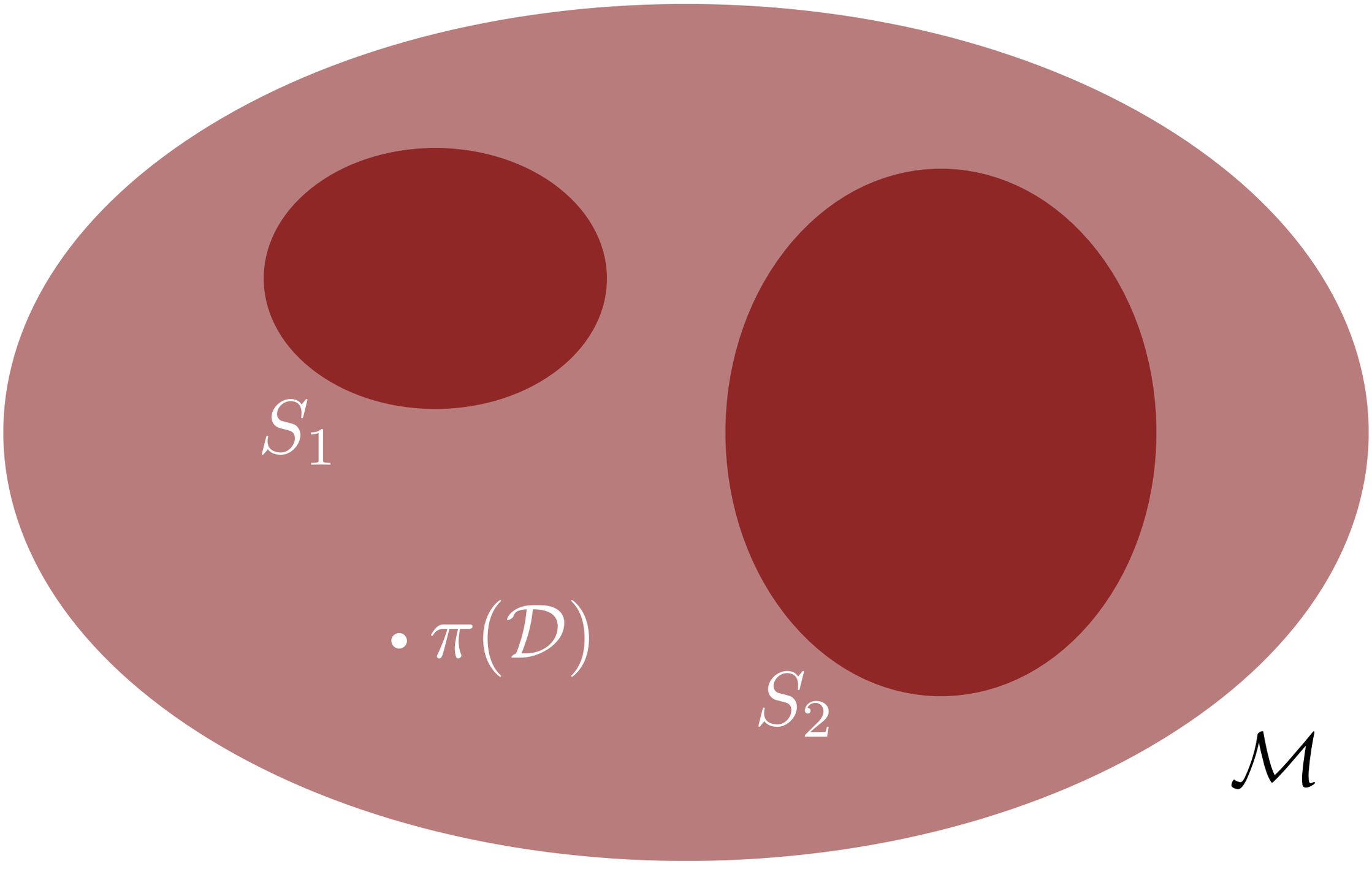
Unfortunately, the marginal likelihood is very difficult to estimate for any nontrivial model.

$$\pi_S(\widetilde{\mathcal{D}}) = \int \mathrm{d}\theta\, \pi_S(\widetilde{\mathcal{D}}|\theta)\pi_S(\theta)$$

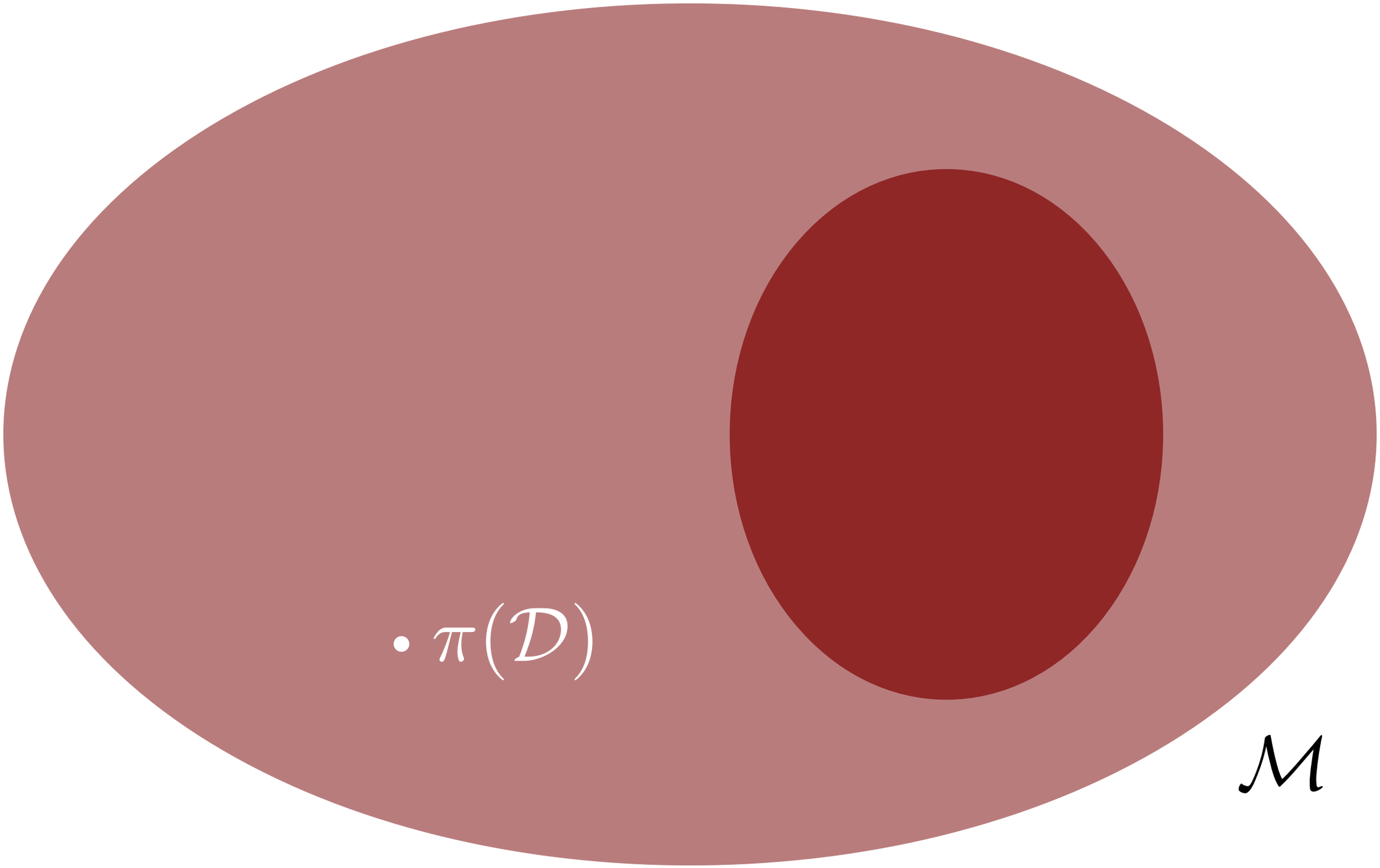
Moreover, these marginal likelihood techniques perform well only within the context of the assumptions.



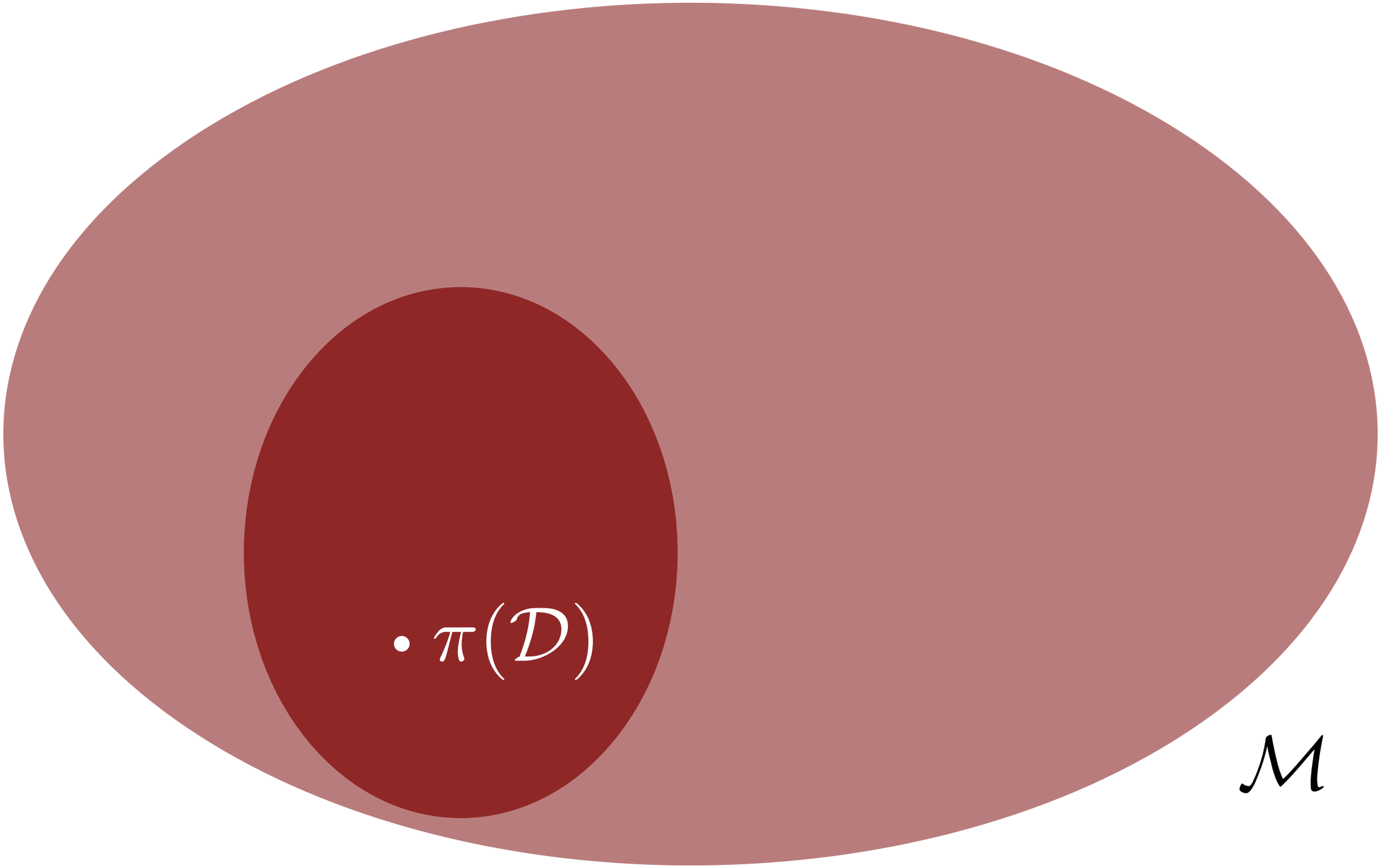
Moreover, these marginal likelihood techniques perform well only within the context of the assumptions.



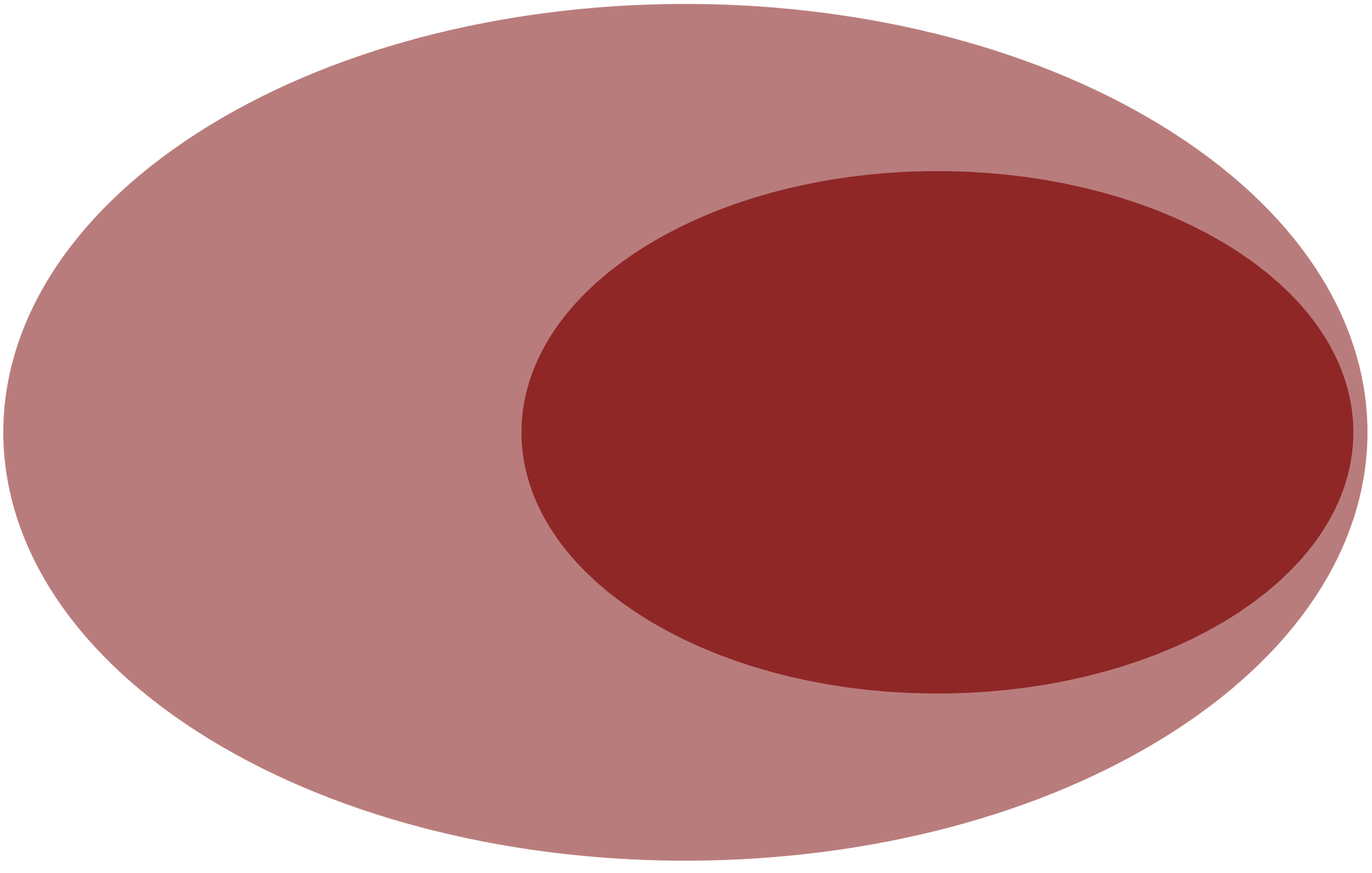
Moreover, these marginal likelihood techniques perform well only within the context of the assumptions.



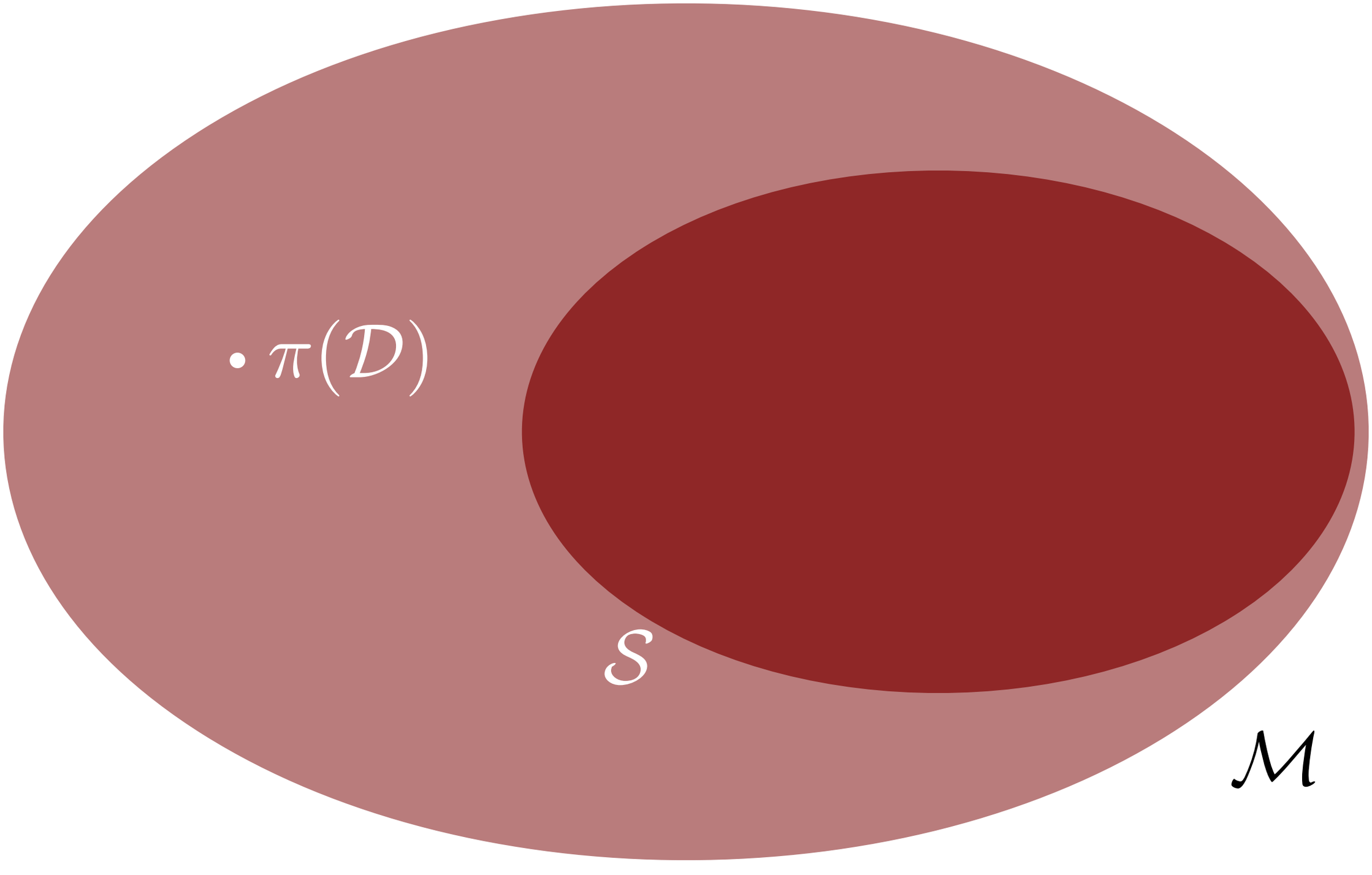
Moreover, these marginal likelihood techniques perform well only within the context of the assumptions.



Predictive Model Comparison



The only way to test the assumptions themselves
is by considering *predictive performance*.



The average data generation process over the entire model is given by the *posterior predictive distribution*.

$$\pi_S(\mathcal{D}|\theta)$$

The average data generation process over the entire model is given by the *posterior predictive distribution*.

$$\pi_S(\mathcal{D}|\widetilde{\mathcal{D}}) = \int \mathrm{d}\theta\, \pi_S(\mathcal{D}|\theta)\pi_S(\theta|\widetilde{\mathcal{D}})$$

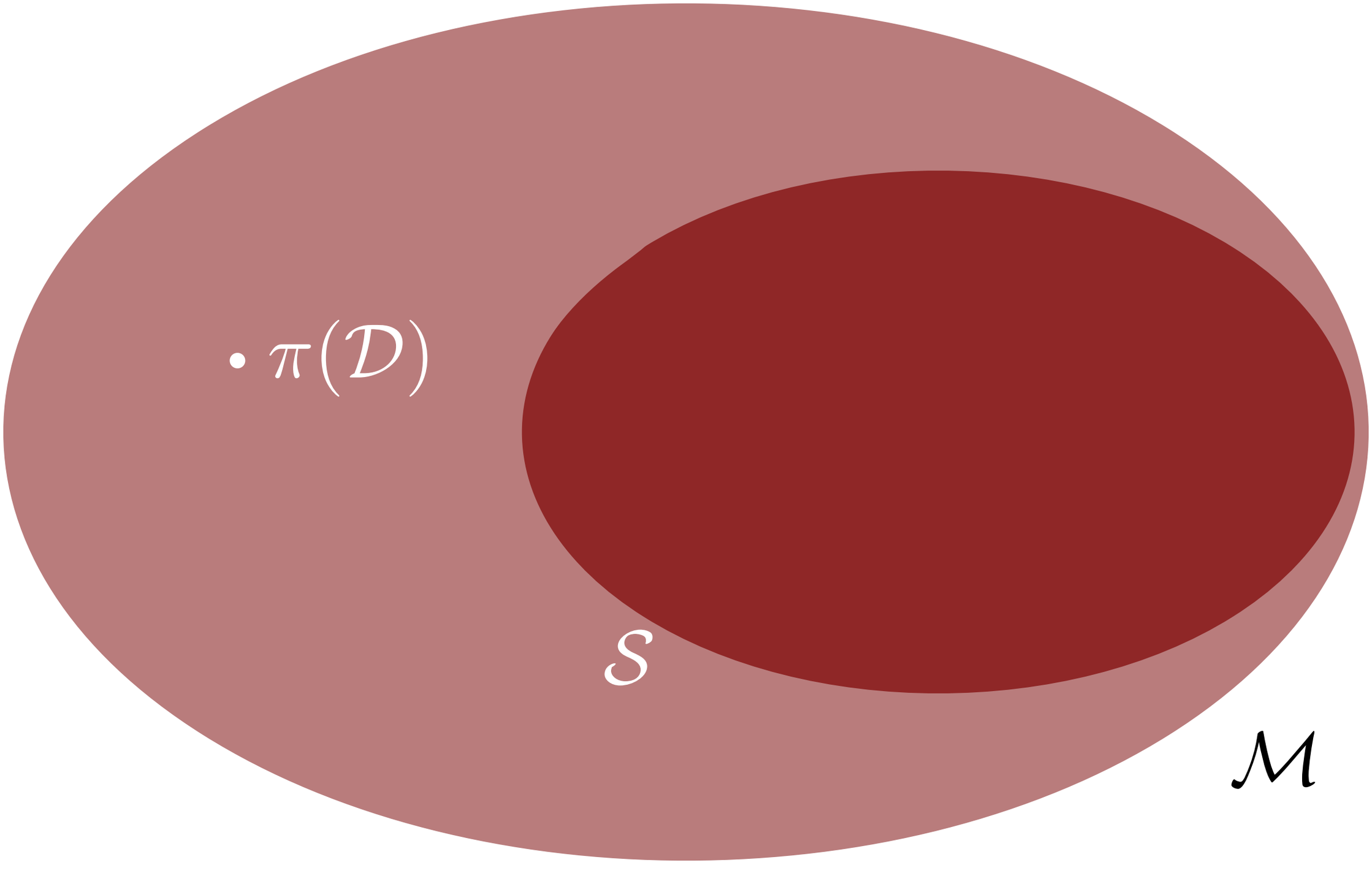
Comparing the posterior predictive distribution to the true data generation process tests our assumptions.

$$\pi_S(\mathcal{D}|\mathcal{D}) \quad \overset{?}{\approx} \quad \pi(\mathcal{D})$$

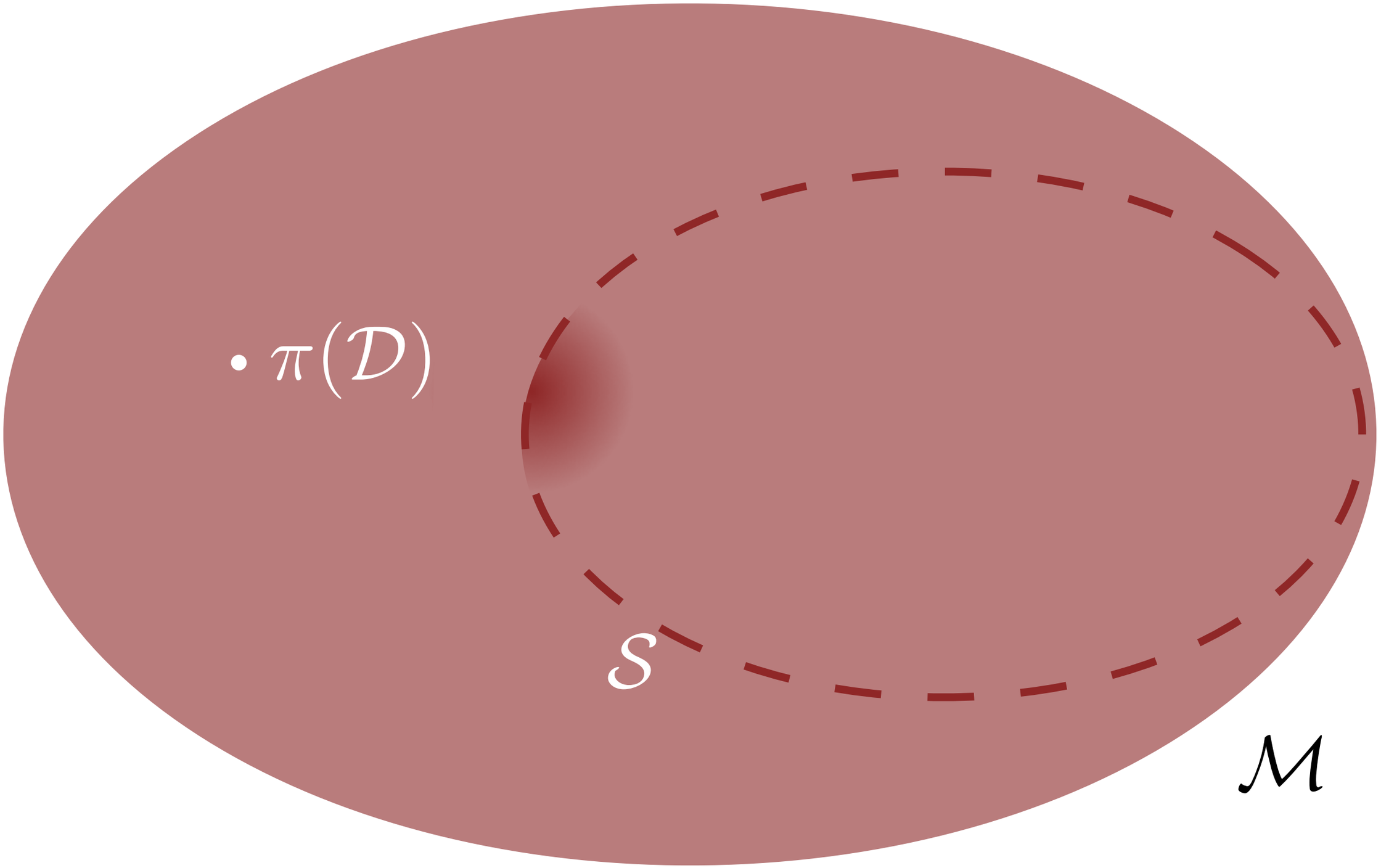
Comparing the posterior predictive distribution to the true data generation process tests our assumptions.

$$\pi_S(\mathcal{D}|\mathcal{D}) \quad \overset{?}{\sim} \quad \mathcal{D}$$

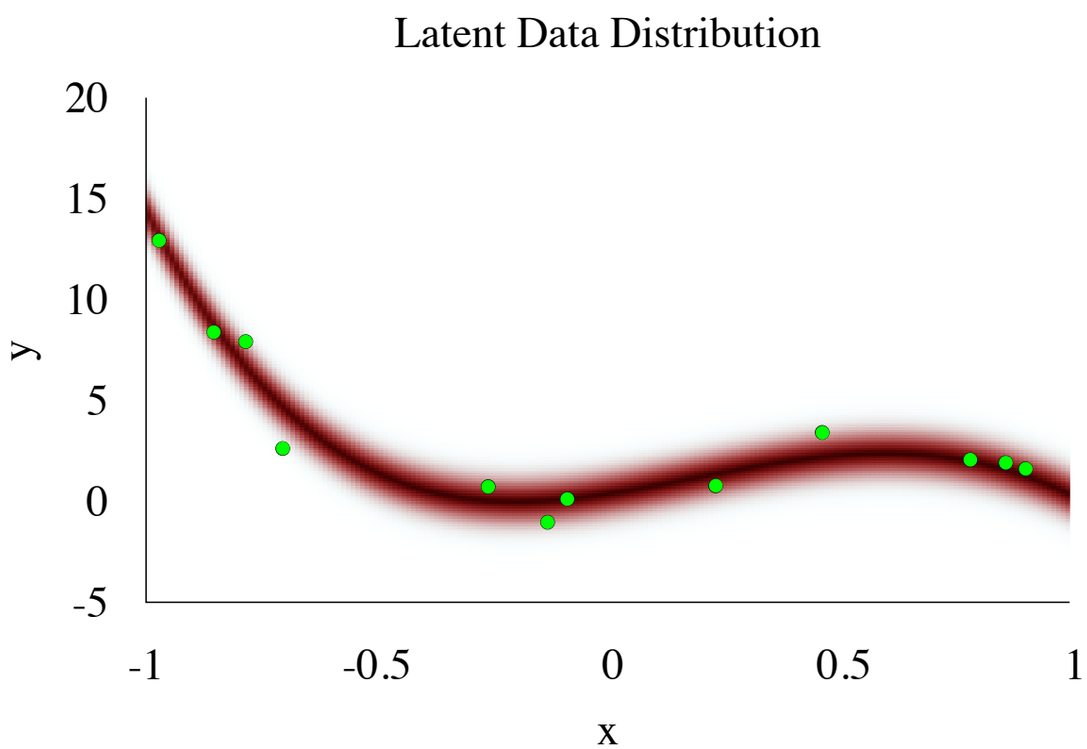
When the small world does not contain the latent data generating process our models will, in general, misfit.

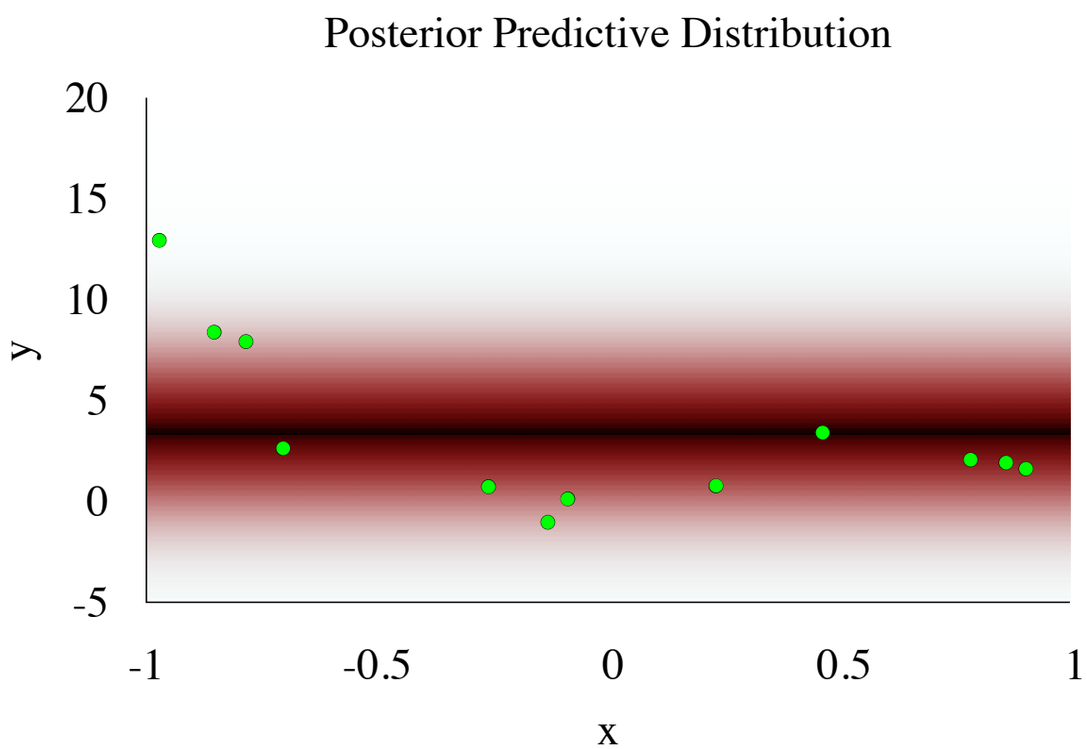


When the small world does not contain the latent data generating process our models will, in general, misfit.

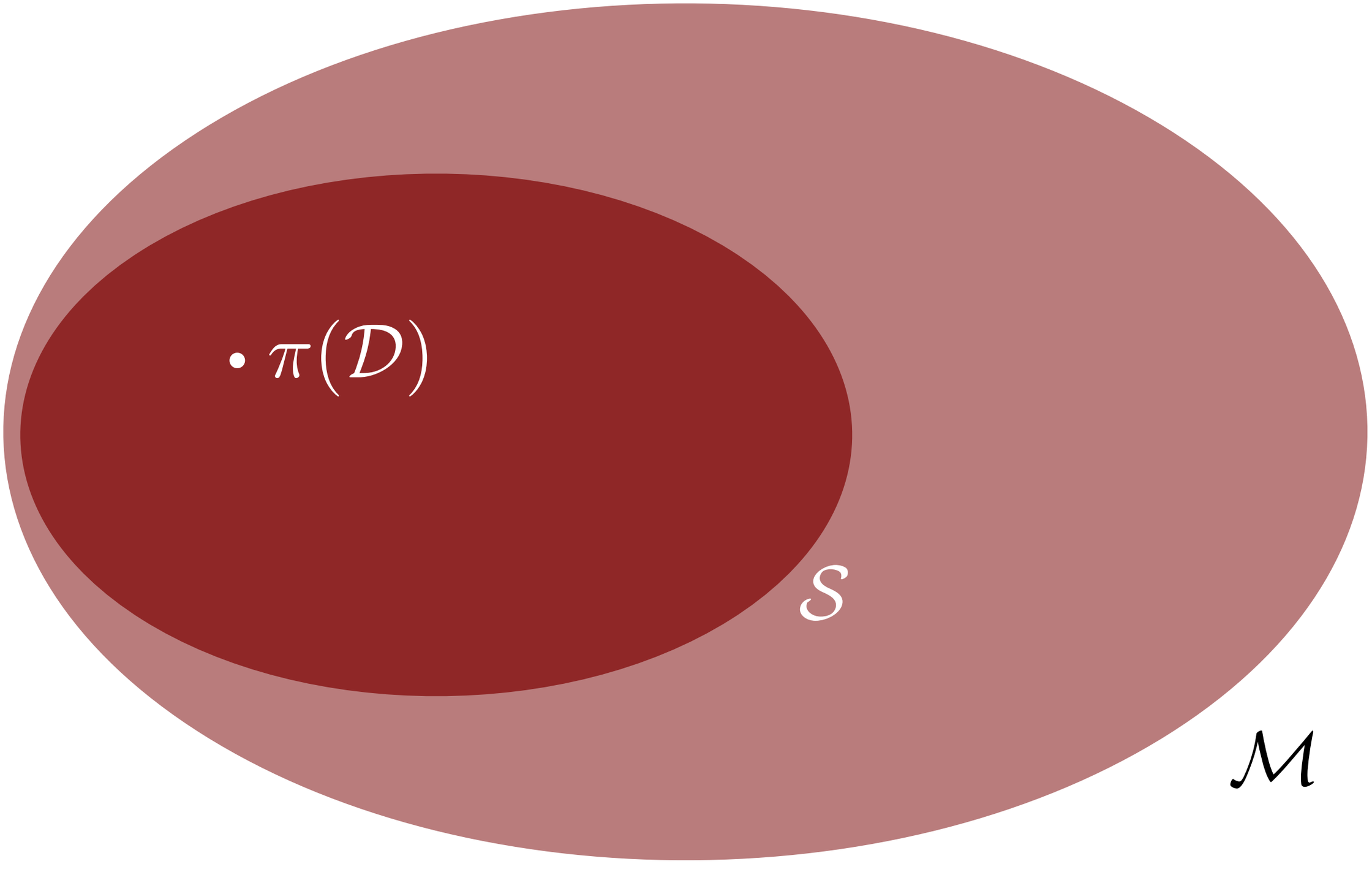


Fortunately, misfit results in tension between predictive distributions and measurements.

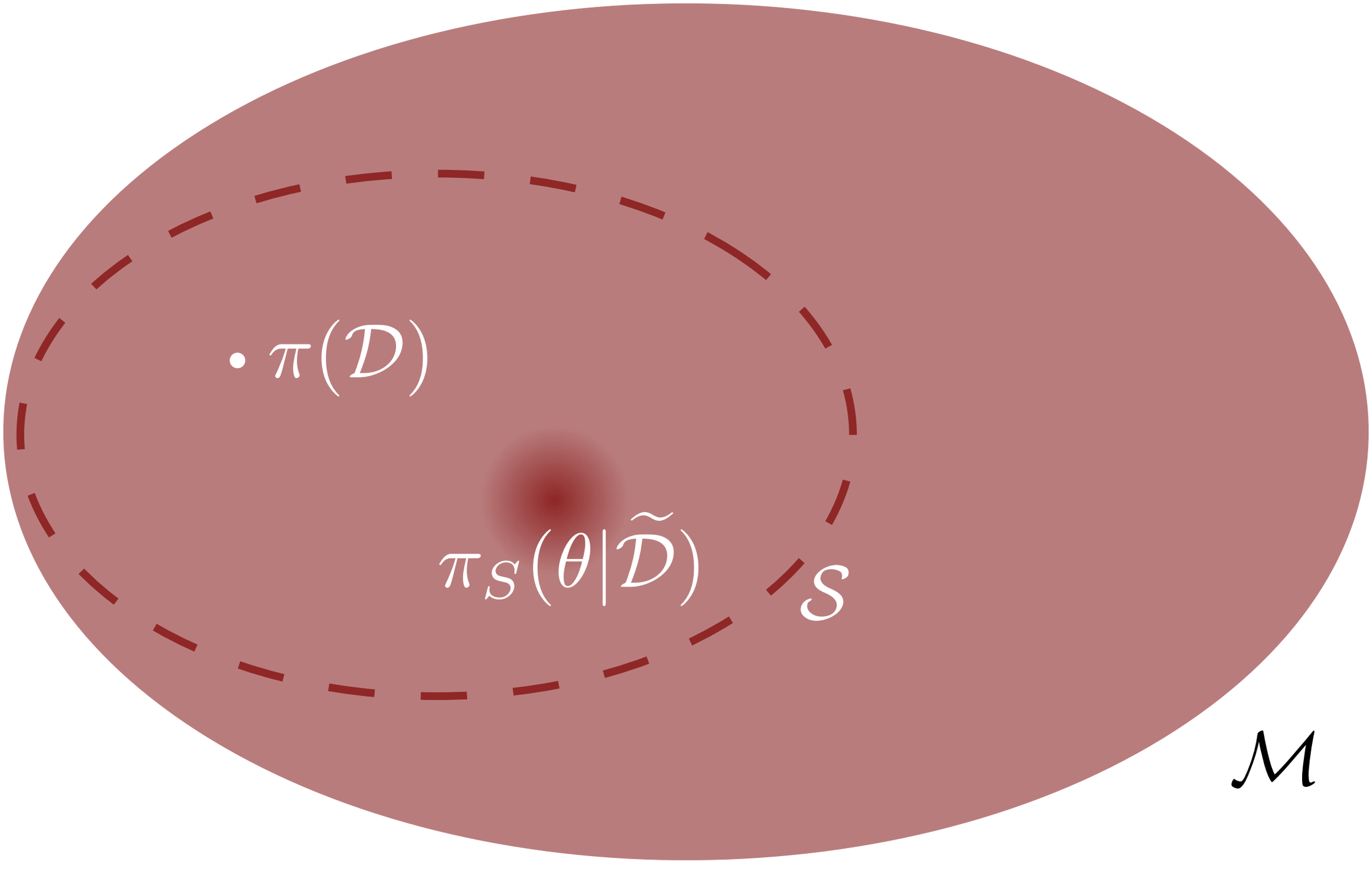




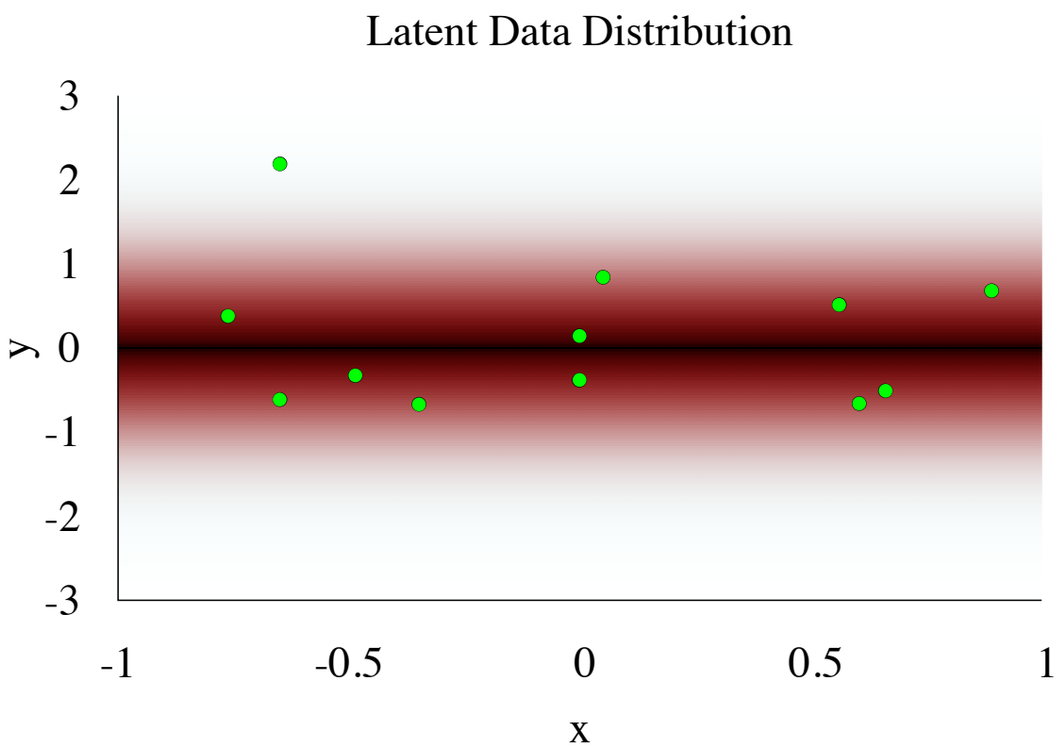
Even if the small world does contain the latent data generating process, however, our models can still overfit.

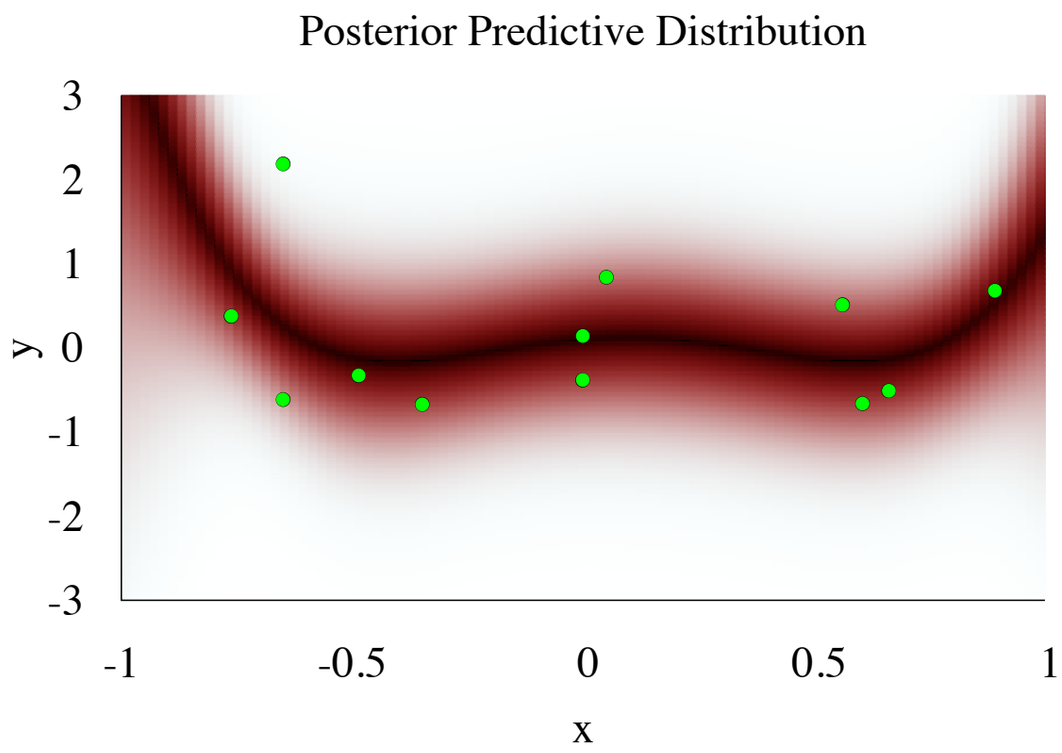


Even if the small world does contain the latent data generating process, however, our models can still overfit.

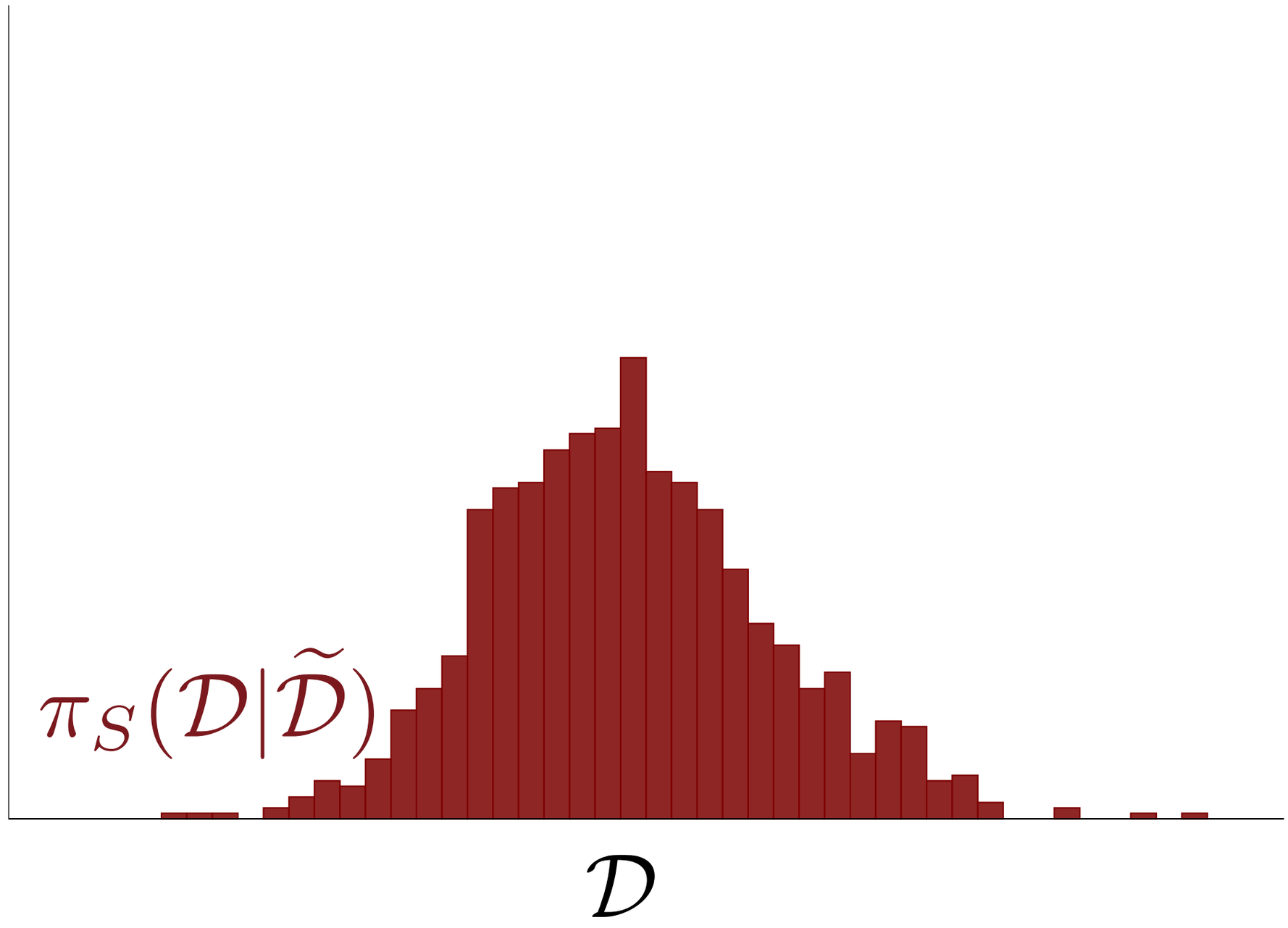


As with misfit, overfitting manifests as tension between predictive distributions and measurements.

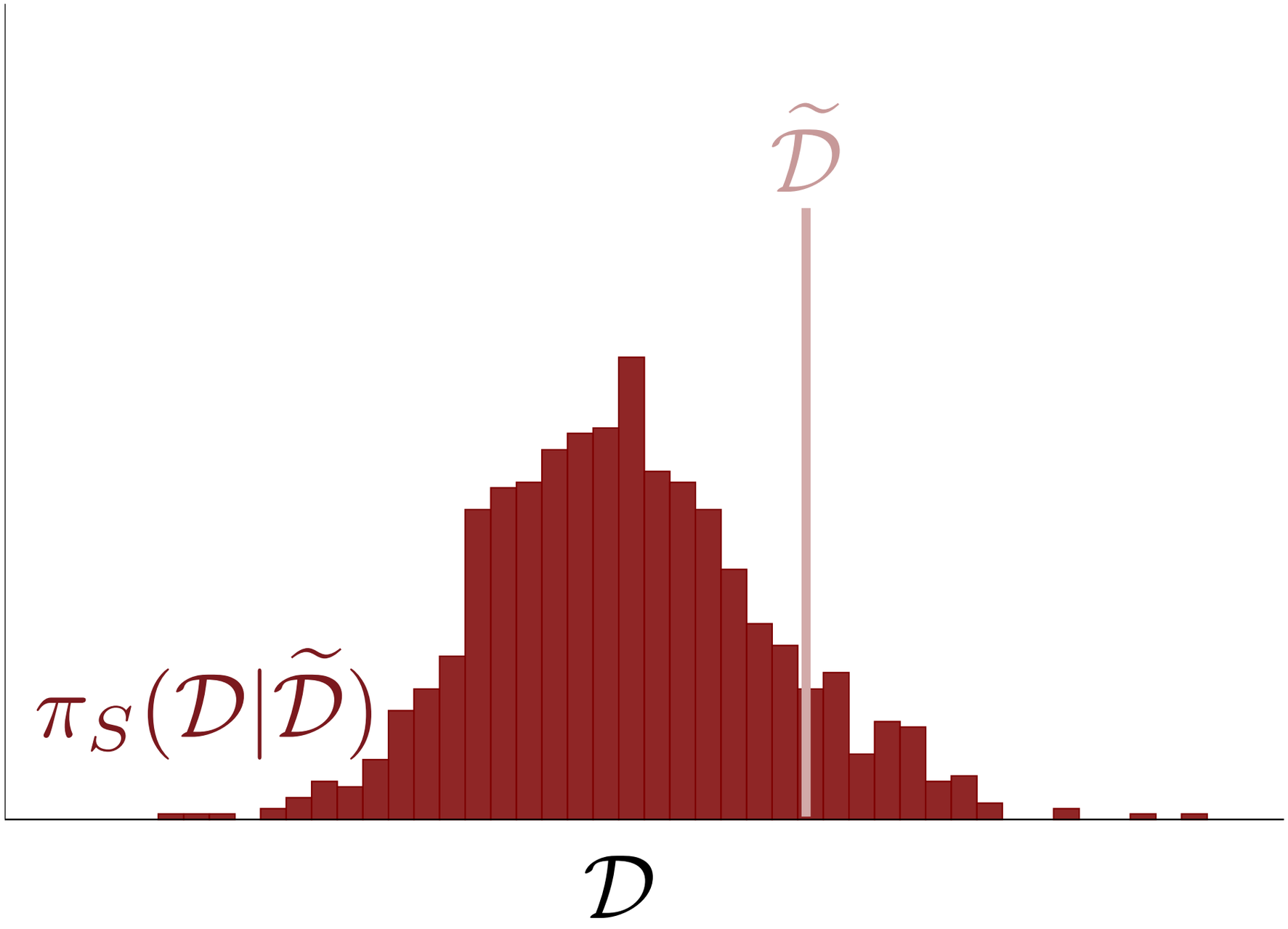




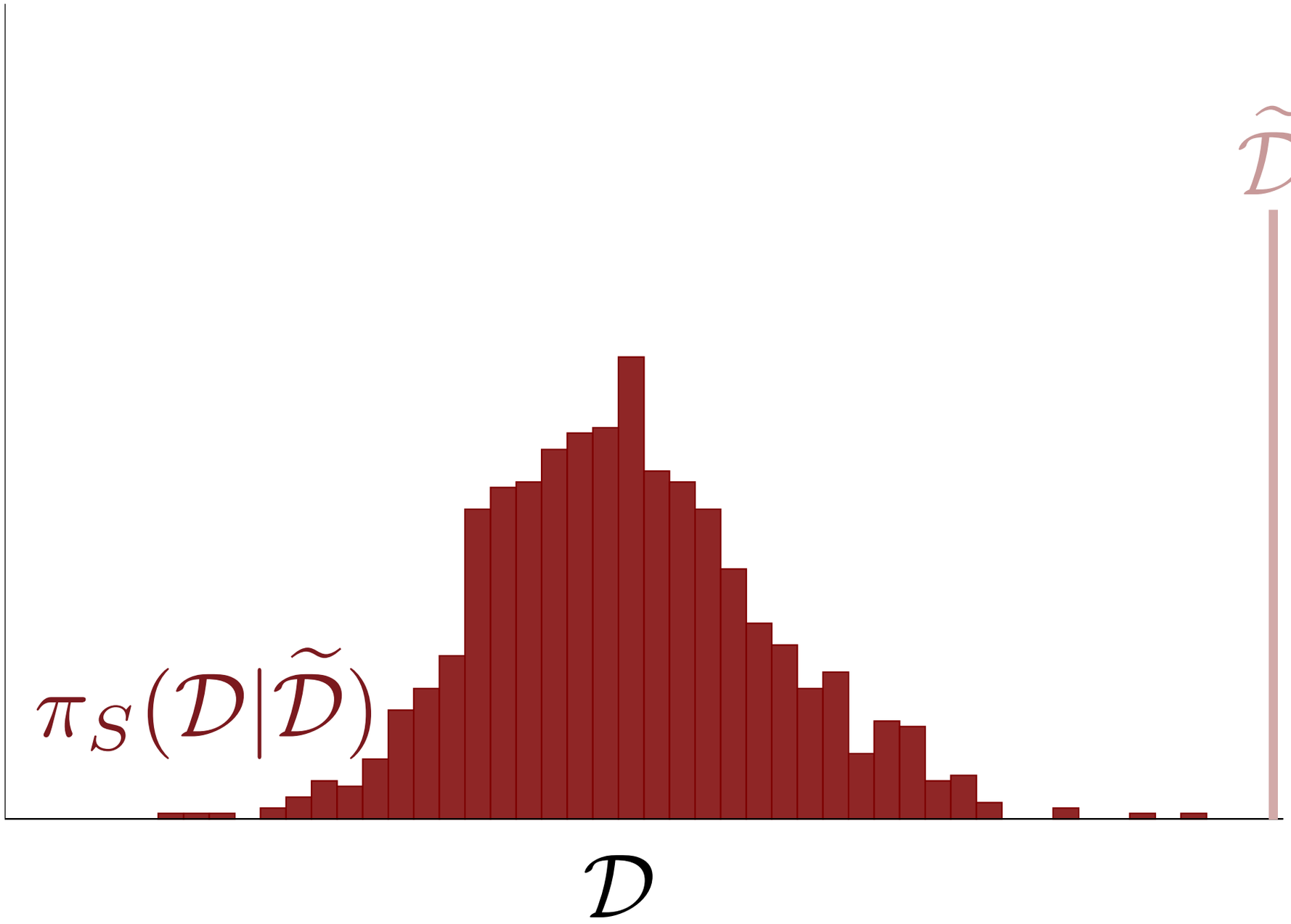
Firstly we can check to see how consistent the measurement is with the inferred predictive distribution.



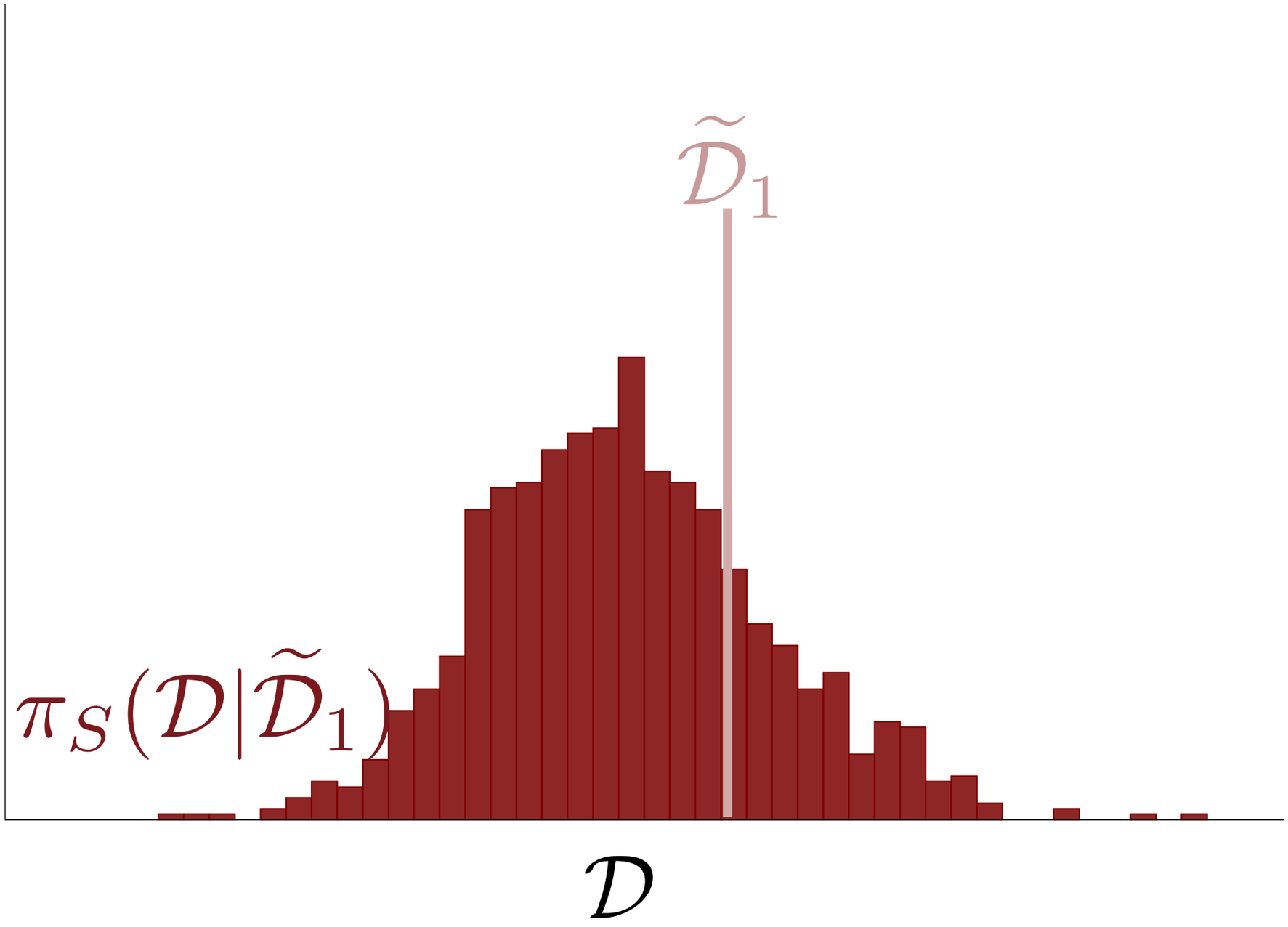
Firstly we can check to see how consistent the measurement is with the inferred predictive distribution.



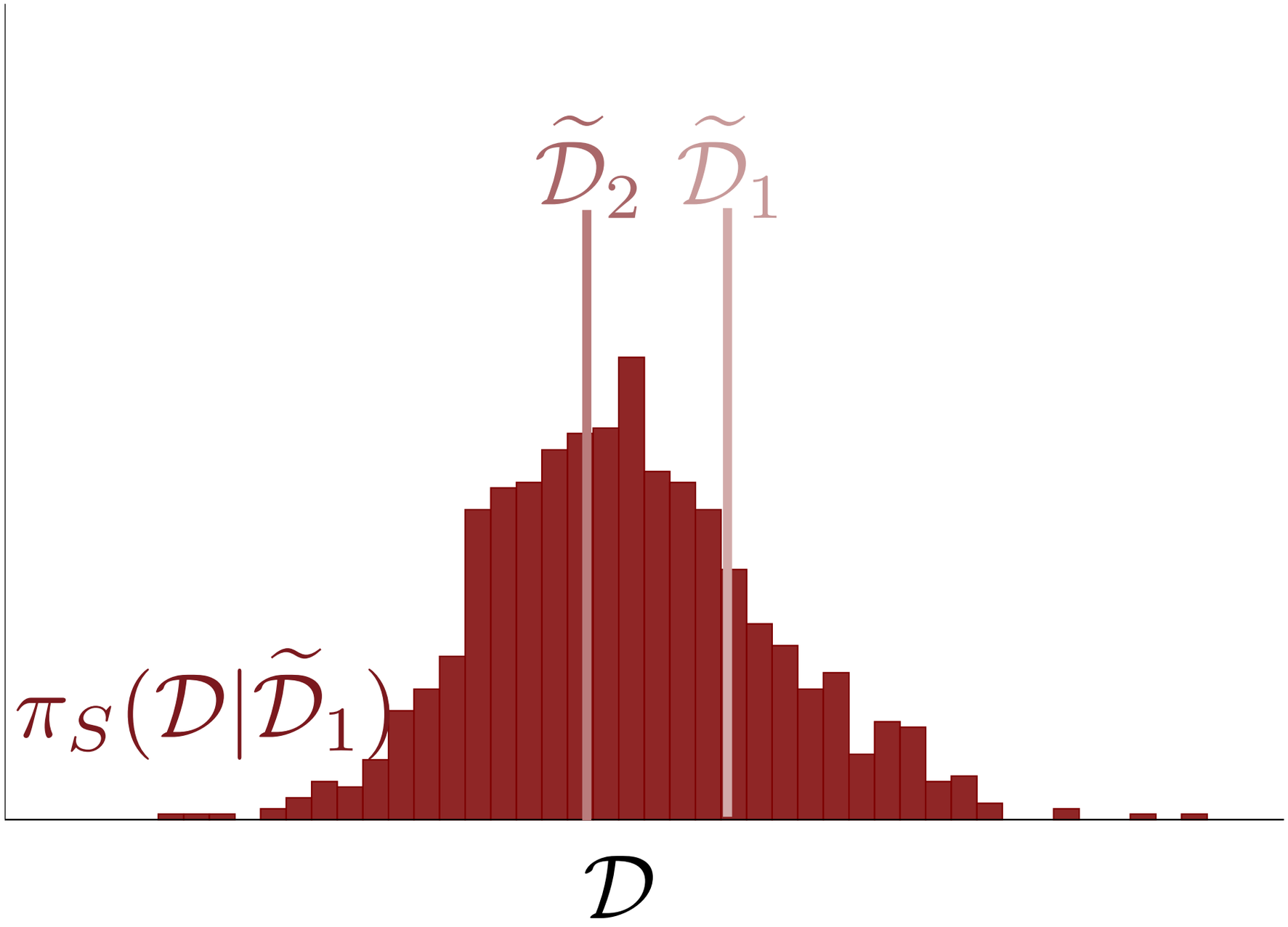
Firstly we can check to see how consistent the measurement is with the inferred predictive distribution.



Similarly, we can check for overfitting by comparing held-out or partitioned measurements.



Similarly, we can check for overfitting by comparing held-out or partitioned measurements.



Similarly, we can check for overfitting by comparing held-out or partitioned measurements.