Rare Event Analysis of Large Language Models

Jake McAllister Dorman, Edward Gillman, Dominic C. Rose, Jamie F. Mair, Juan P. Garrahan

University of Nottingham, United Kingdom

Machine Learning in Science @ Nottingham

Authors

Jake McAllister Dorman

Edward Gillman

Dominic C. Rose

Jamie F. Mair

Juan P. Garrahan

Collaborators

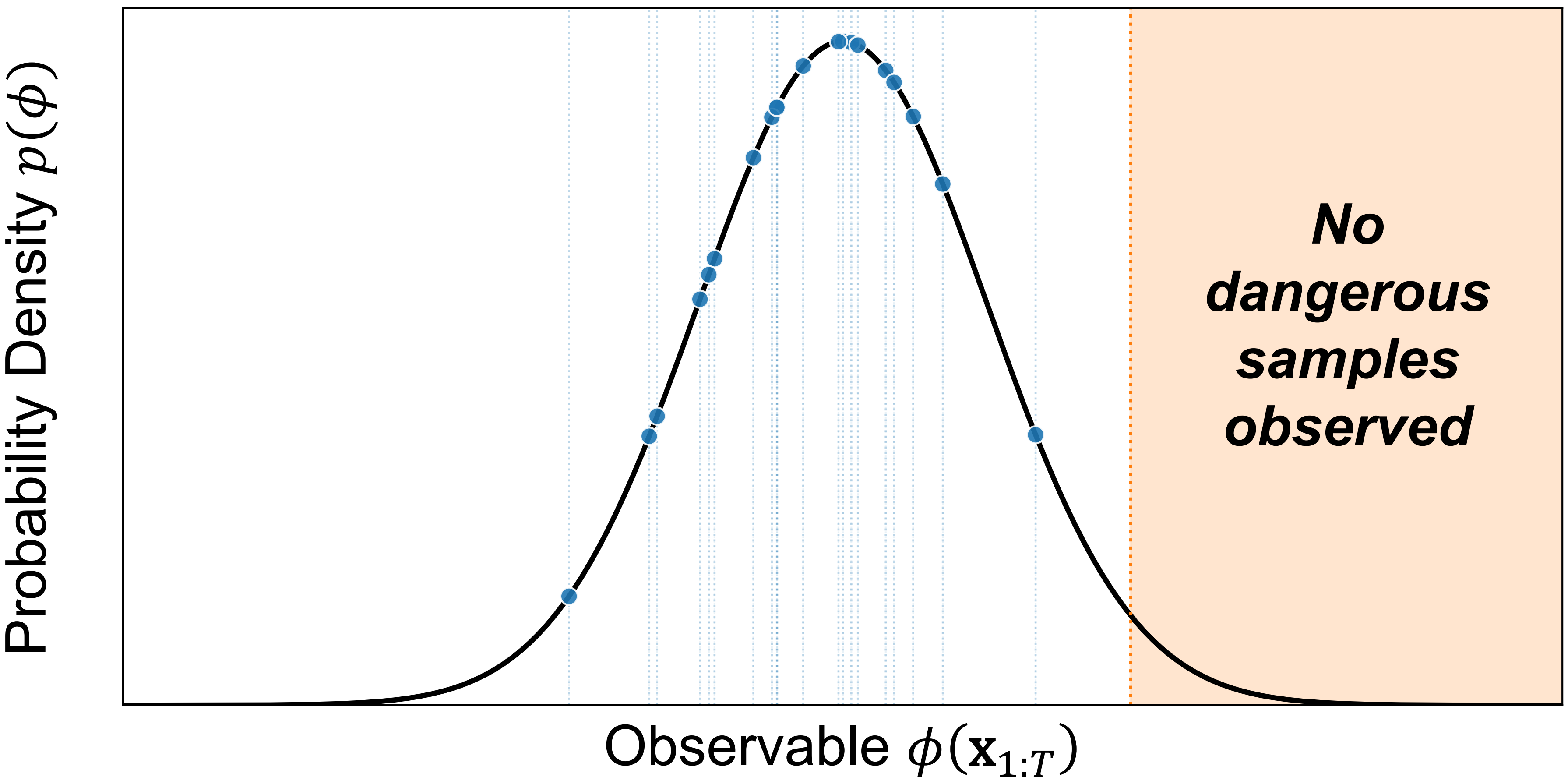
Josiah Oliver-Dee

Tyler A. Rolfe

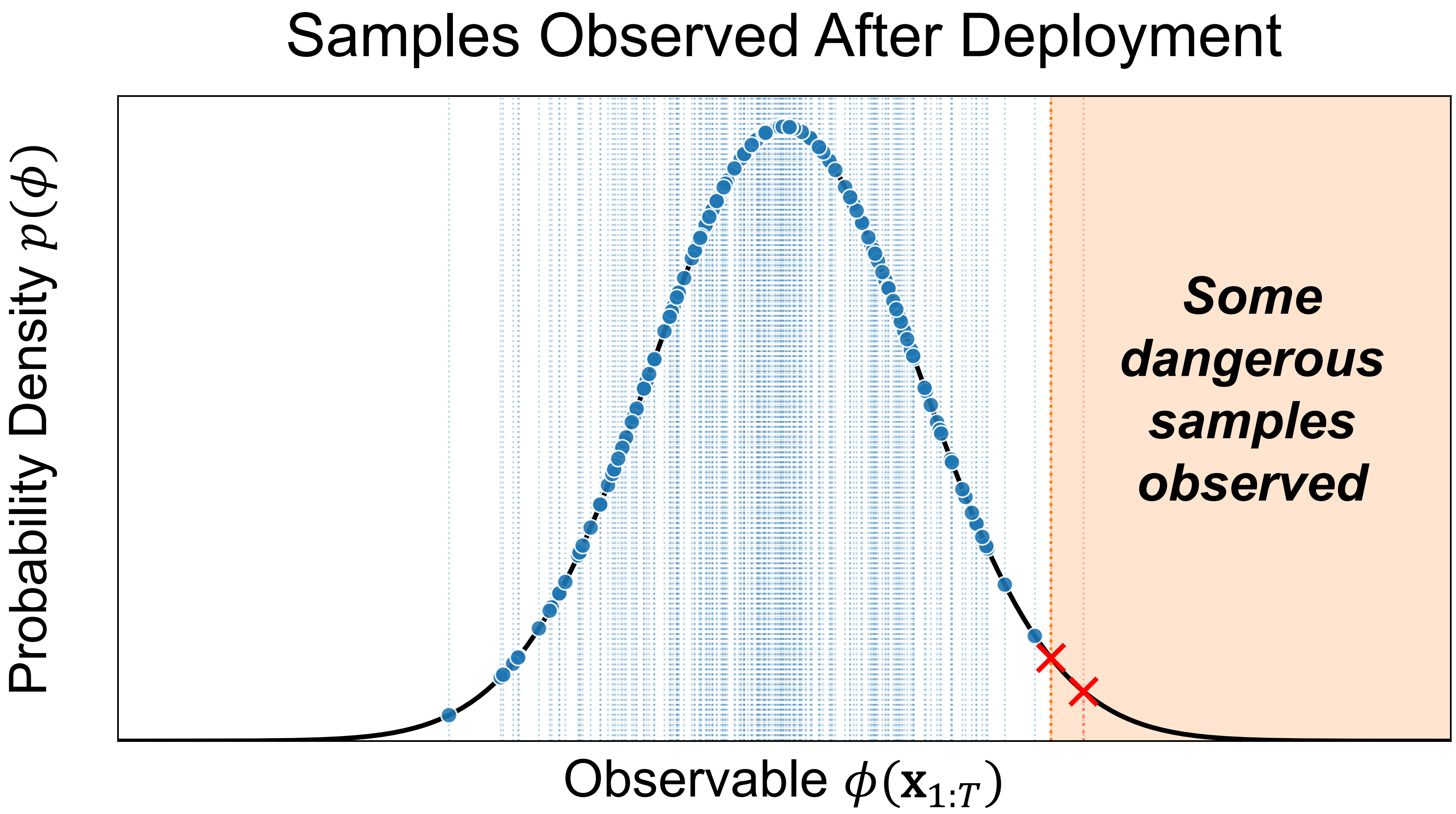
Alvin Vinod Chandran

Rare Events in Large Language Models

Samples Observed During Testing



Rare Events in Large Language Models



Contents

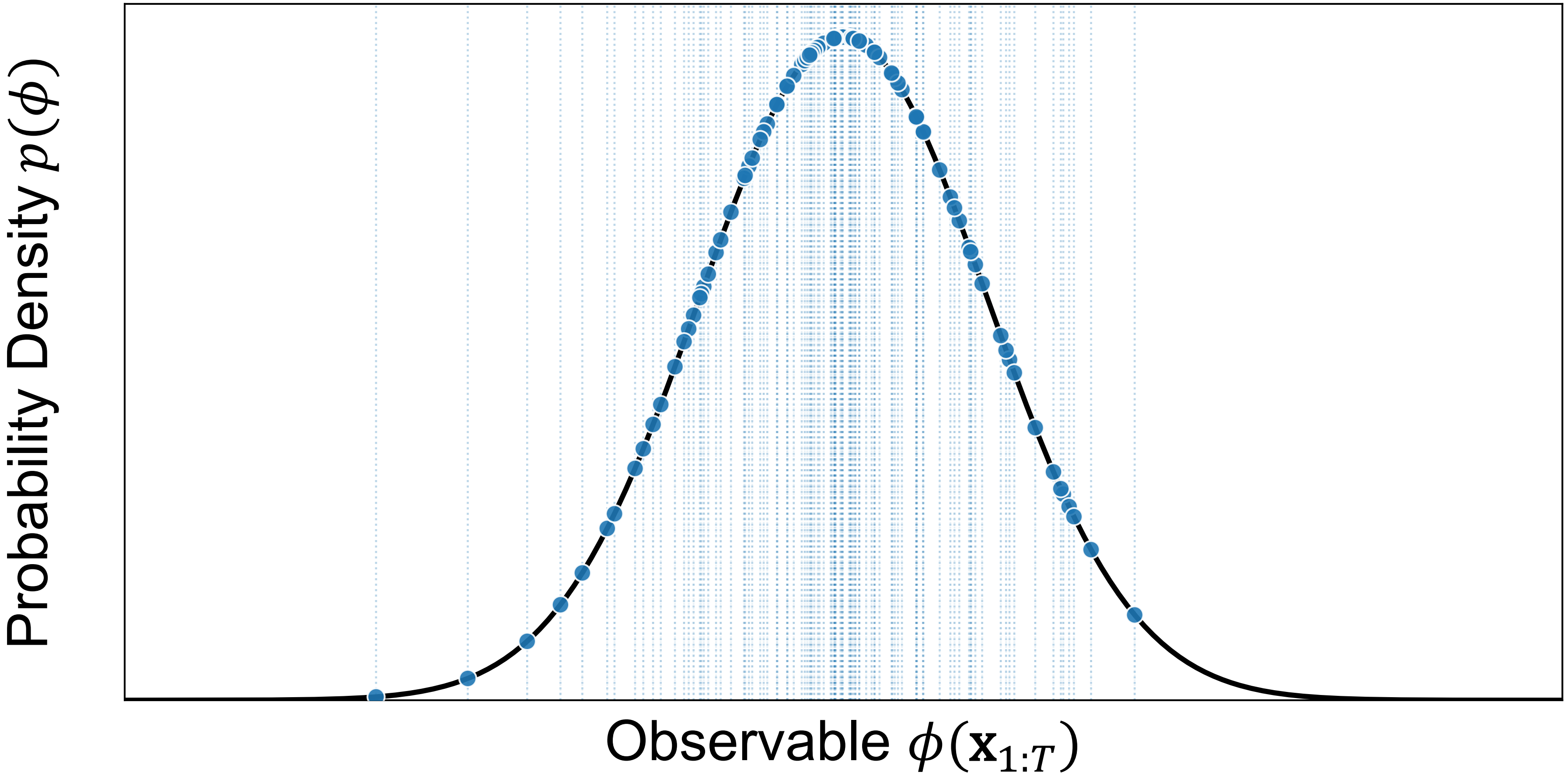
We present a framework to efficiently sample and estimate the likelihood of these rare, unwanted completions.

We use an MCMC approach using methods from Statistical Physics and Computational Chemistry to sample these rare events.

Our method enabled a $\mathbf{10^8}$ **times improvement** over direct sampling.

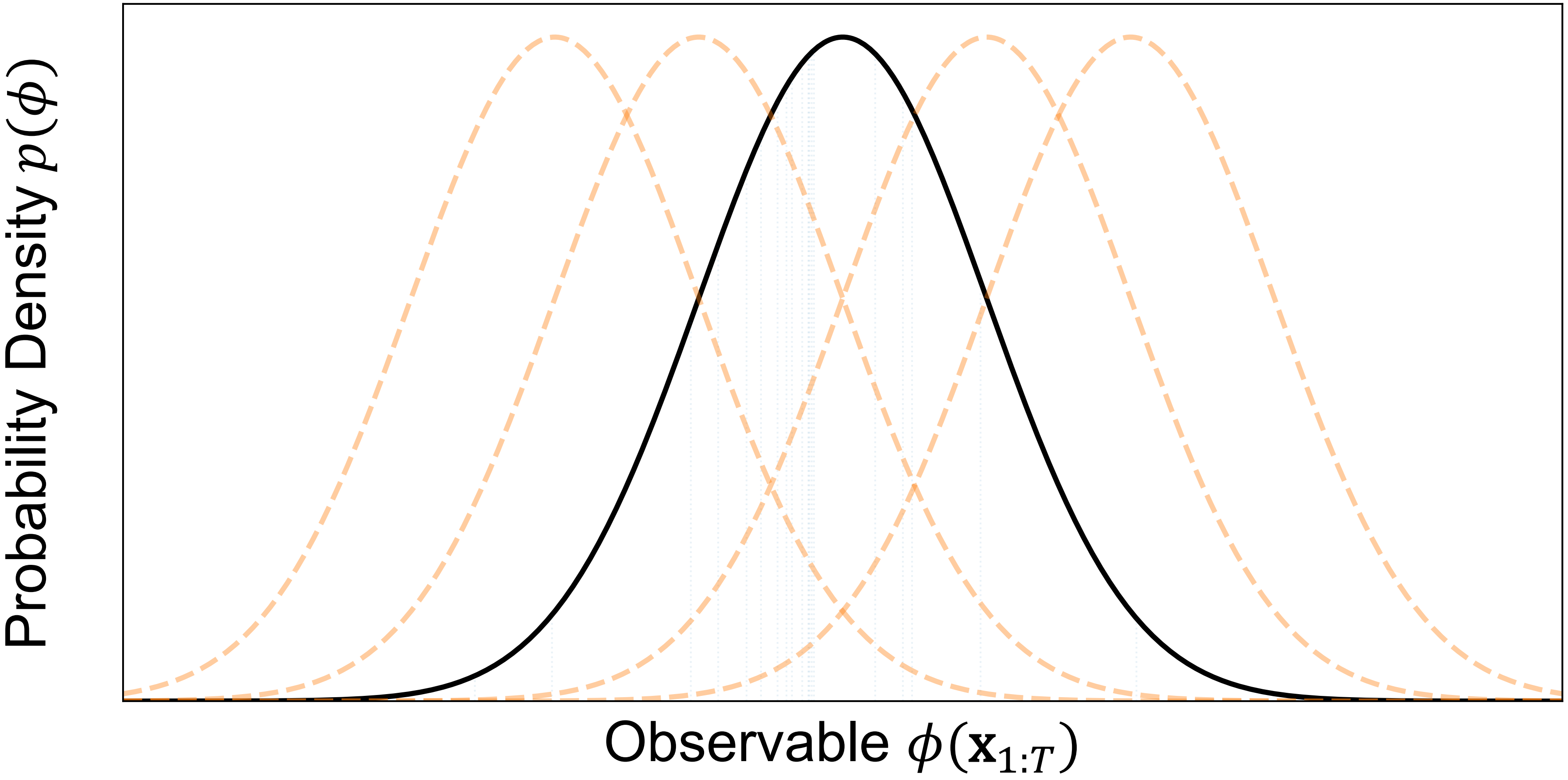
Importance Sampling

Sampling the model directly is unlikely to give rare samples…



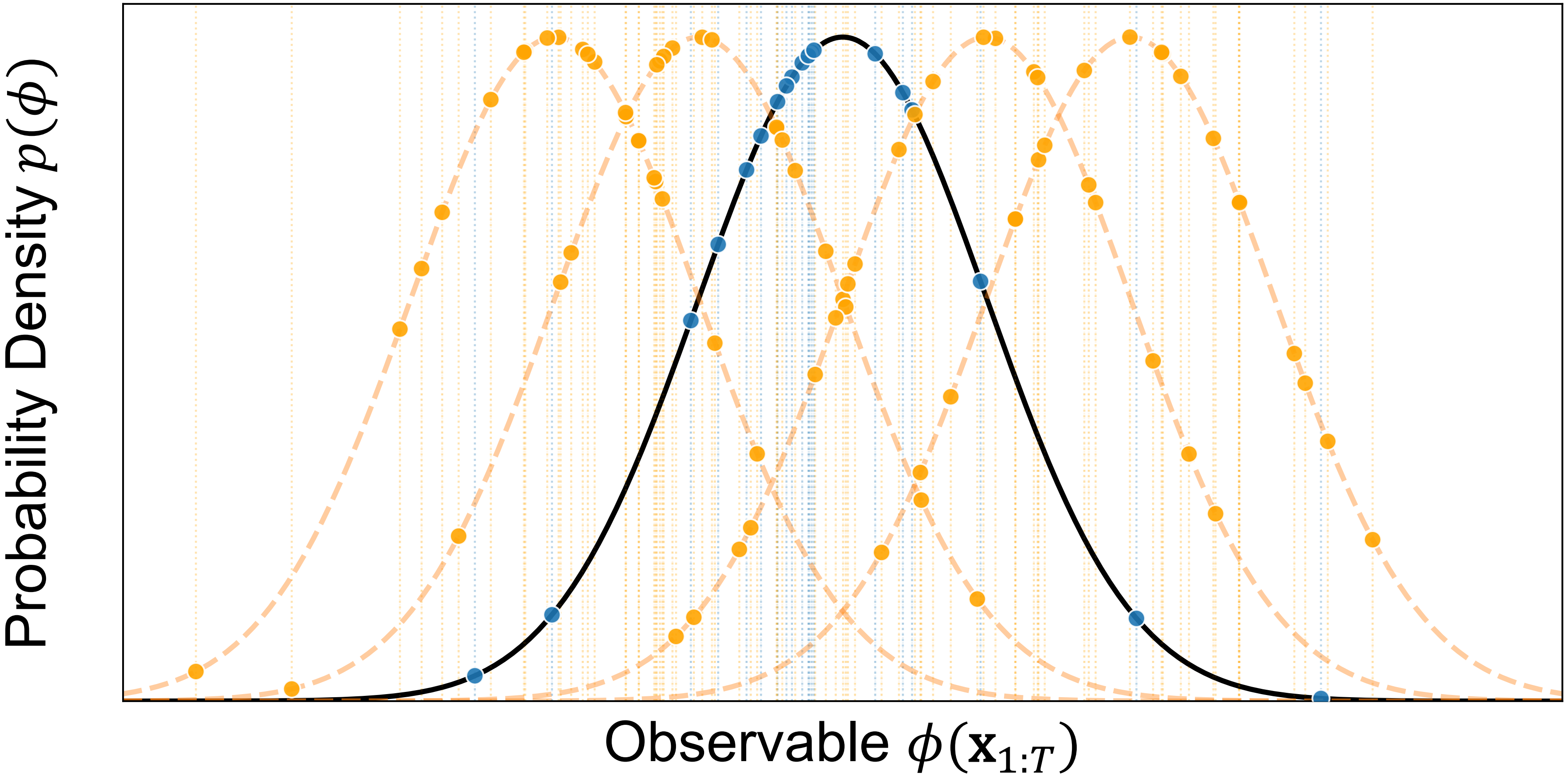
Importance Sampling

Instead, construct other distributions with better coverage.



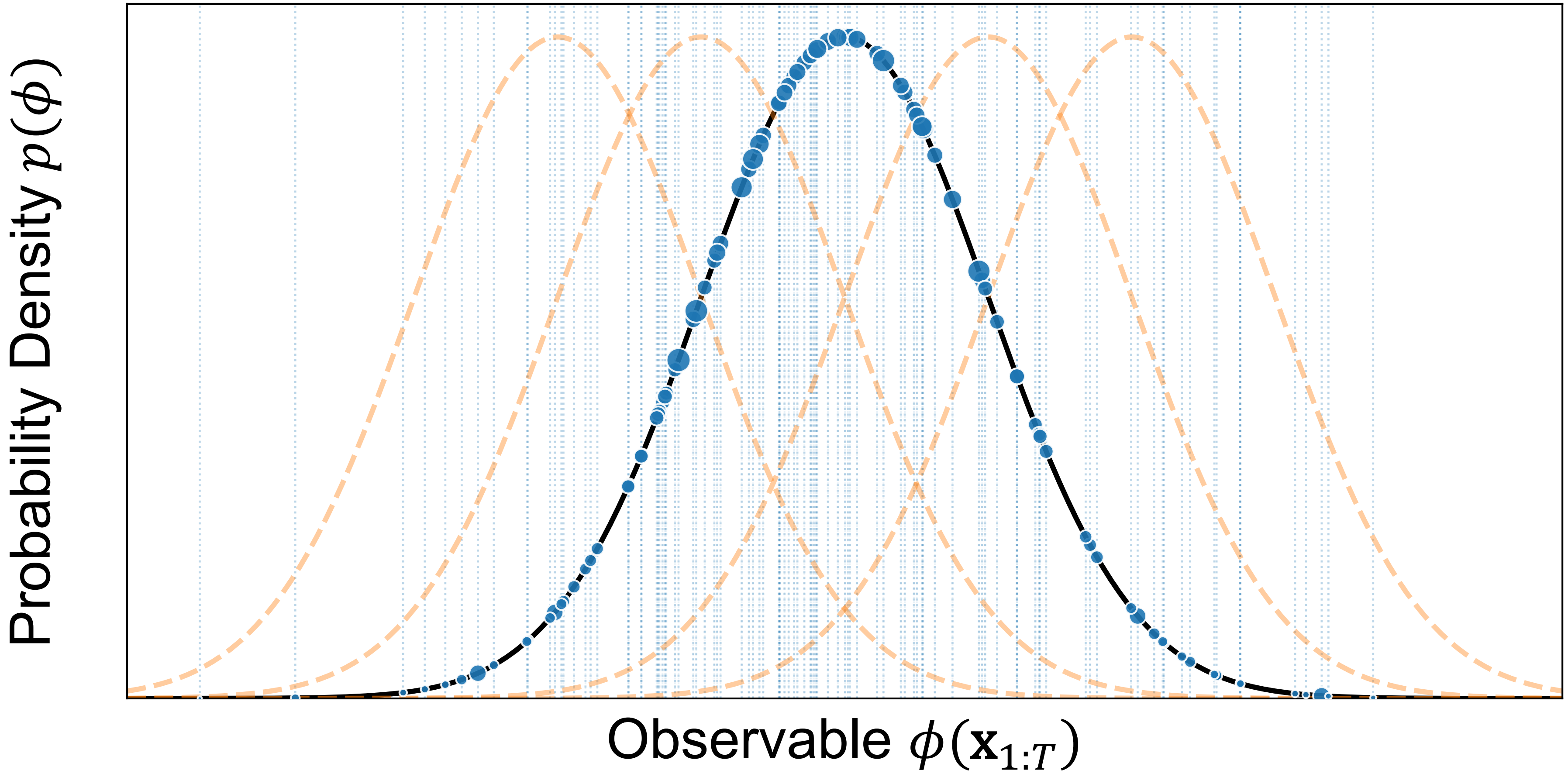
Importance Sampling

We can then sample these…



Importance Sampling

…and reweight these samples back to original distribution.



Exponentially Reweighted Distributions

We want to sample the 'closest' distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

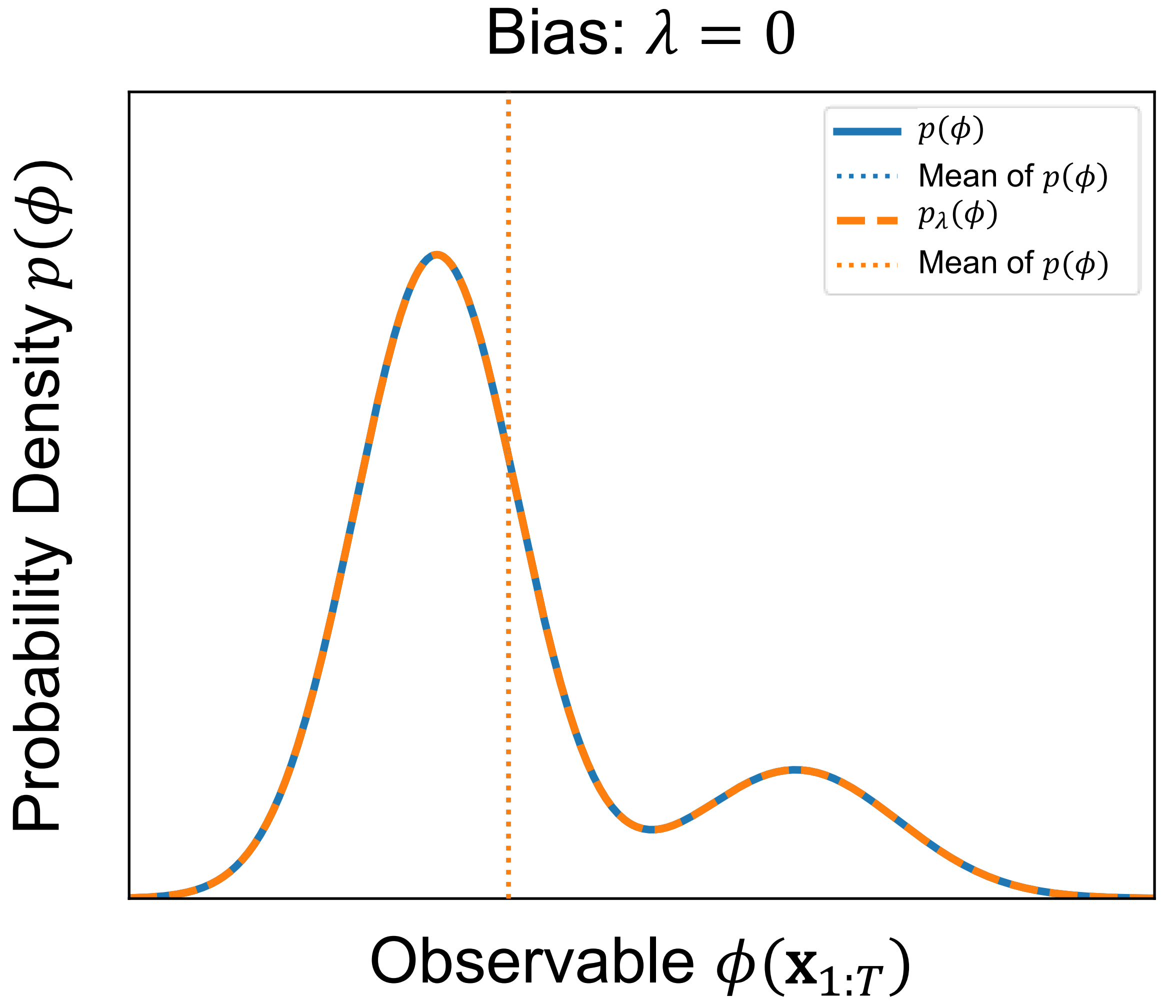
By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.

Exponentially Reweighted Distributions

We want to sample the ‘closest’ distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.

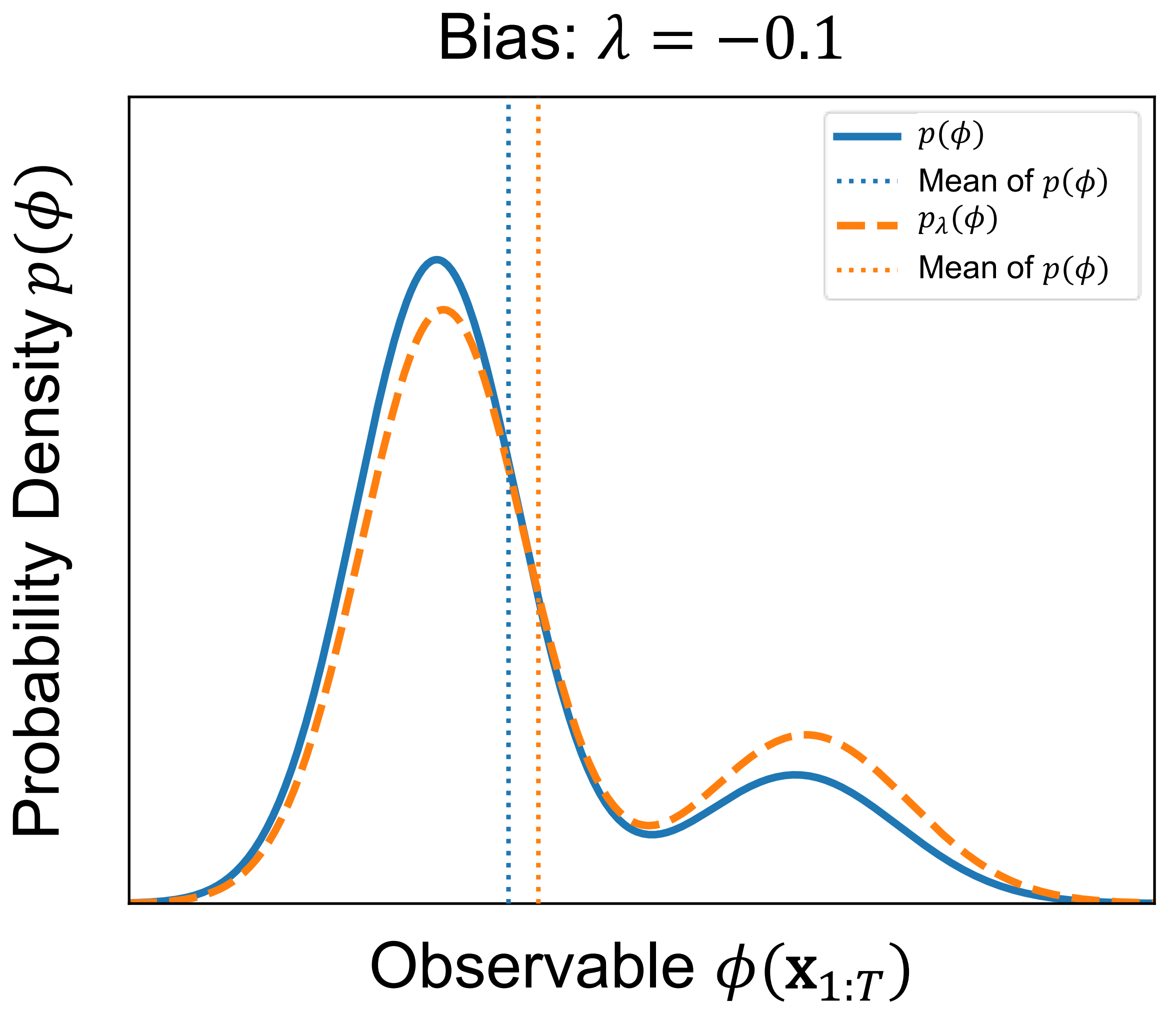


Exponentially Reweighted Distributions

We want to sample the 'closest' distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.

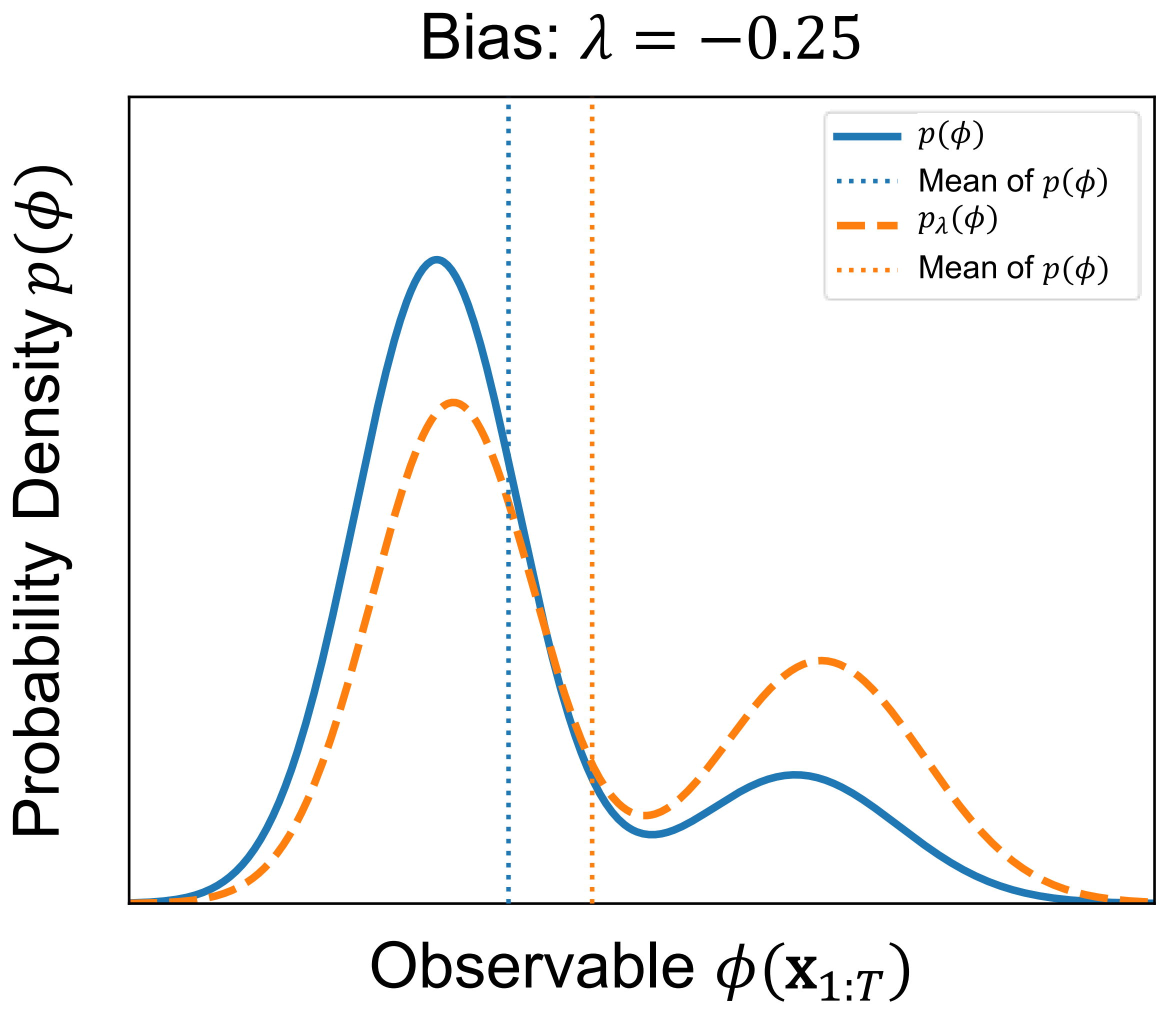


Exponentially Reweighted Distributions

We want to sample the 'closest' distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.



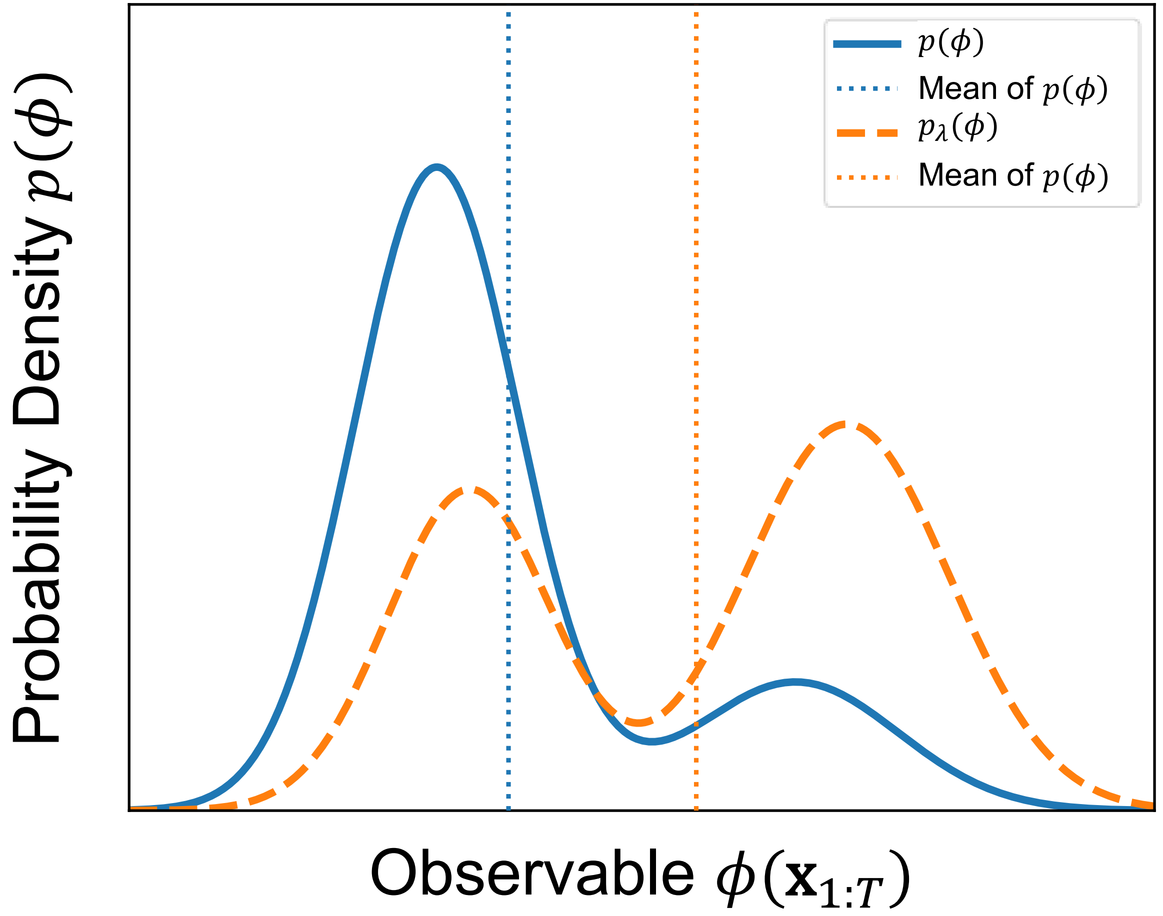
Exponentially Reweighted Distributions

We want to sample the 'closest' distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.

Bias: $\lambda = -0.5$

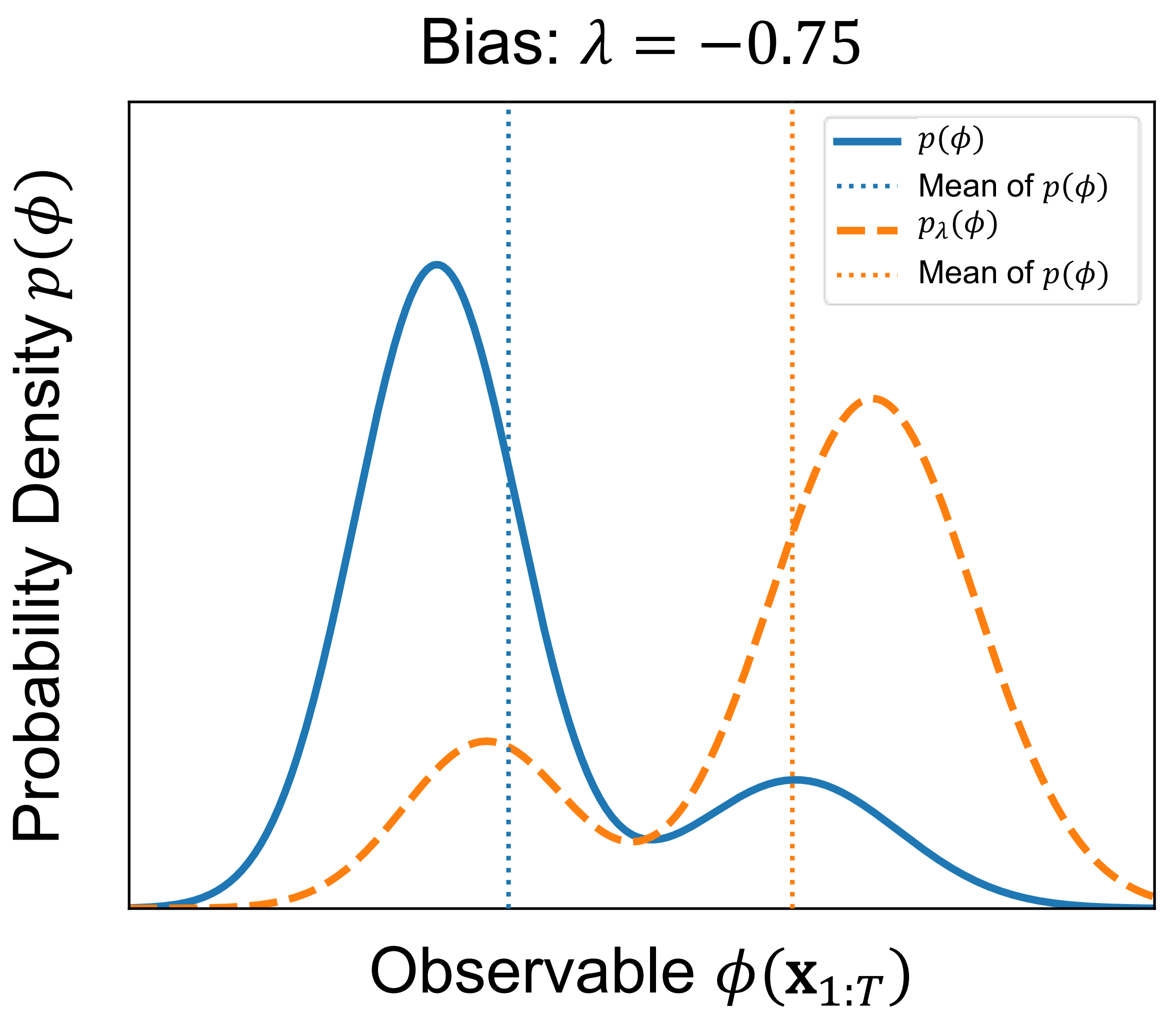


Exponentially Reweighted Distributions

We want to sample the 'closest' distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.

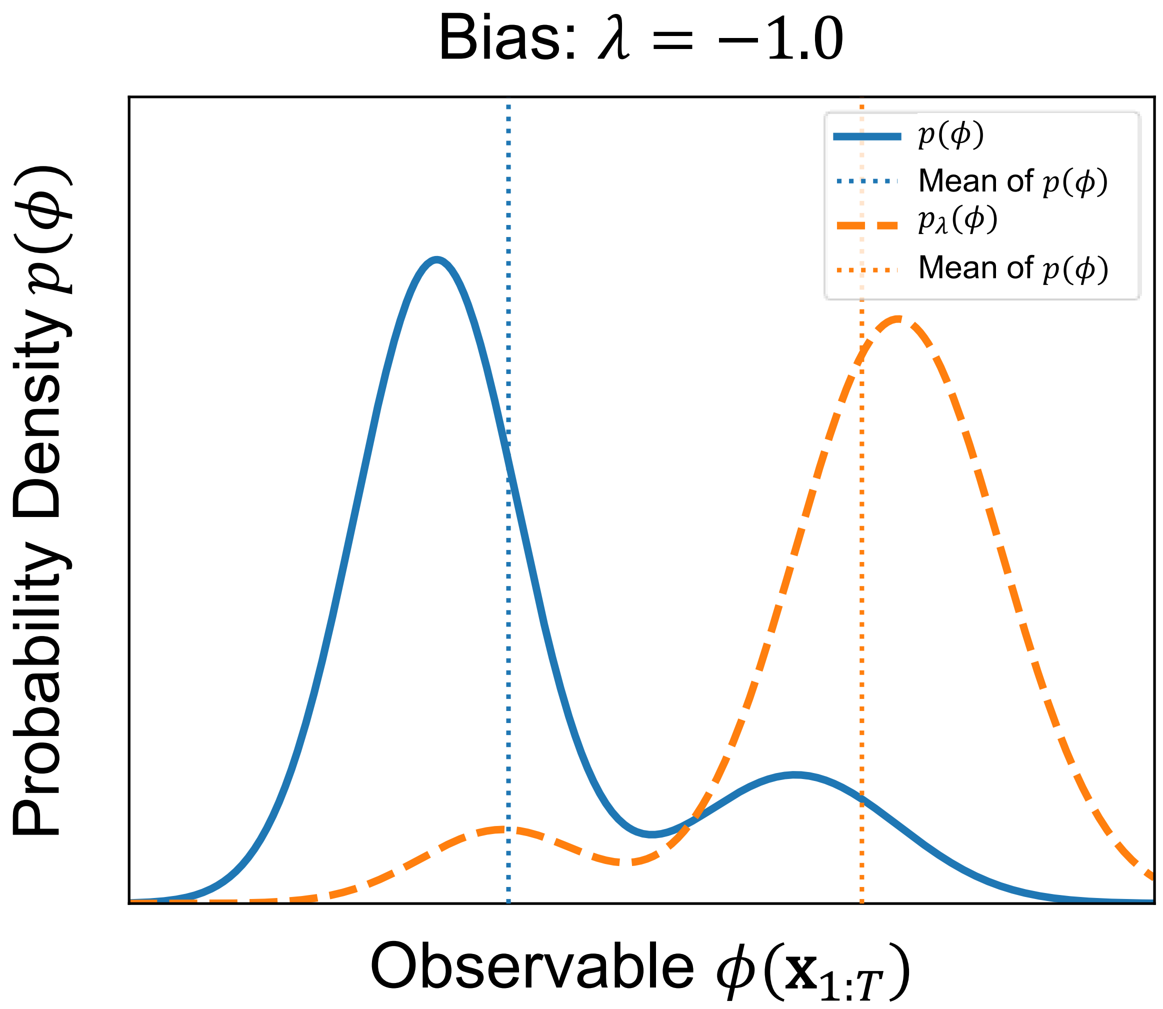


Exponentially Reweighted Distributions

We want to sample the 'closest' distributions to our model, with rare values of our observable $\phi(\mathbf{x})$. We choose the *exponential reweighting*:

$$p_\lambda(\mathbf{x}_{1:T}) = \frac{1}{Z(\lambda)} e^{-\lambda\phi(\mathbf{x}_{1:T})} p(\mathbf{x}_{1:T}).$$

By changing the *bias* λ, we shift the mean of $\phi(\mathbf{x}_{1:T})$ towards rarer values.



Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \text{ARI}$):

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi =$ ARI):

Step	*Prompt* $\mathbf{x}_0$	Completion $\mathbf{x}_{1:T}$
Step s:	$\mathbf{x}_{1:T}^{s} =$ "*The sky is* blue and cloudy. What day is it?"	

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \text{ARI}$):

Step *Prompt* $\mathbf{x}_0$ Completion $\mathbf{x}_{1:T}$

Step s: $\mathbf{x}_{1:T}^{s}=$ "*The sky is* blue and cloudy. What day is it?"

Proposal: $\mathbf{x}_{1:T'}^{*}=$ "*The sky is* blue, radiant and dazzling."

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \text{ARI}$):

Step	*Prompt* $\mathbf{x}_0$ Completion $\mathbf{x}_{1:T}$	$\text{ARI}(\mathbf{x}_{1:T})$
Step s:	$\mathbf{x}_{1:T}^{s} =$ "*The sky is* blue and cloudy. What day is it?"	-3.86
Proposal:	$\mathbf{x}_{1:T'}^{*} =$ "*The sky is* blue, radiant and dazzling."	2.26

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = $ ARI):

Step	*Prompt* $\mathbf{x}_0$	Completion $\mathbf{x}_{1:T}$	$\mathrm{ARI}(\mathbf{x}_{1:T})$	Acceptance
Step s:	$\mathbf{x}^s_{1:T} =$ "*The sky is*	blue and cloudy. What day is it?"	-3.86	1.0
Proposal:	$\mathbf{x}^*_{1:T'} =$ "*The sky is*	blue, radiant and dazzling."	2.26	

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = $ ARI):

Step	Prompt $\mathbf{x}_0$ / Completion $\mathbf{x}_{1:T}$	$\mathrm{ARI}(\mathbf{x}_{1:T})$	Acceptance	$\epsilon \sim U[0,1]$
Step s:	$\mathbf{x}^s_{1:T} =$ "*The sky is* blue and cloudy. What day is it?"	-3.86	1.0	0.37
Proposal:	$\mathbf{x}^*_{1:T'} =$ "*The sky is* blue, radiant and dazzling."	2.26		

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi =$ ARI):

Step	*Prompt* $\mathbf{x}_0$	Completion $\mathbf{x}_{1:T}$

Step s+1: $\mathbf{x}_{1:T'}^{s+1} =$ "*The sky is* blue, radiant and dazzling."

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi =$ ARI):

Step *Prompt* $\mathbf{x}_0$ Completion $\mathbf{x}_{1:T}$

Step s+1: $\mathbf{x}_{1:T'}^{s+1} =$ "*The sky is* blue, radiant and dazzling."

Proposal: $\mathbf{x}_{1:T''}^{*} =$ "*The sky is* blue, radiant and cloudy."

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \text{ARI}$):

Step	*Prompt* $\mathbf{x}_0$	Completion $\mathbf{x}_{1:T}$	$\text{ARI}(\mathbf{x}_{1:T})$
Step s+1: $\mathbf{x}^{s+1}_{1:T'}$ =	"*The sky is*	blue, radiant and dazzling."	2.26
Proposal: $\mathbf{x}^{*}_{1:T''}$ =	"*The sky is*	blue, radiant and cloudy."	0.91

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \text{ARI}$):

Step	*Prompt* $\mathbf{x}_0$	Completion $\mathbf{x}_{1:T}$	$\text{ARI}(\mathbf{x}_{1:T})$	Acceptance
Step s+1: $\mathbf{x}_{1:T'}^{s+1} =$	"*The sky is*	blue, radiant and dazzling."	2.26	0.87
Proposal: $\mathbf{x}_{1:T''}^{*} =$	"*The sky is*	blue, radiant and cloudy."	0.91	

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \mathrm{ARI}$):

Step	*Prompt* $\mathbf{x}_0$	Completion $\mathbf{x}_{1:T}$	$\mathrm{ARI}(\mathbf{x}_{1:T})$	Acceptance	$\epsilon \sim U[0,1]$
Step s+1: $\mathbf{x}_{1:T'}^{s+1} =$	"*The sky is*	blue, radiant and dazzling."	2.26	0.87	0.94
Proposal: $\mathbf{x}_{1:T''}^{*} =$	"*The sky is*	blue, radiant and cloudy."	0.91		

Transition Path Sampling

We cannot sample $p_\lambda(\mathbf{x}_{1:T})$ directly as we don't know $Z(\lambda)$.

Instead, we use a variant of Markov Chain Monte Carlo called Transition Path Sampling.

If we have a prompt "The sky is" and we bias towards complicated text ($\phi = \text{ARI}$):

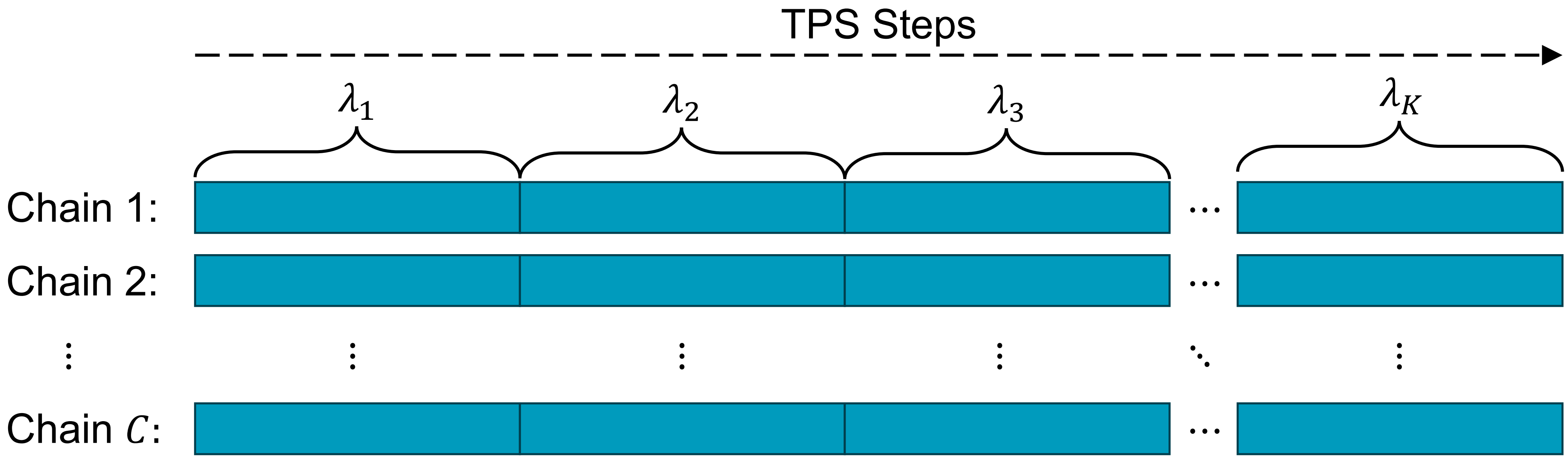
Step | *Prompt* $\mathbf{x}_0$ | Completion $\mathbf{x}_{1:T}$

Step s+2: $\mathbf{x}_{1:T''}^{s+2} =$ "*The sky is* blue, radiant and dazzling."

Annealing & Removing Unconverged Samples

We use annealing, which means varying λ as the chain progresses.

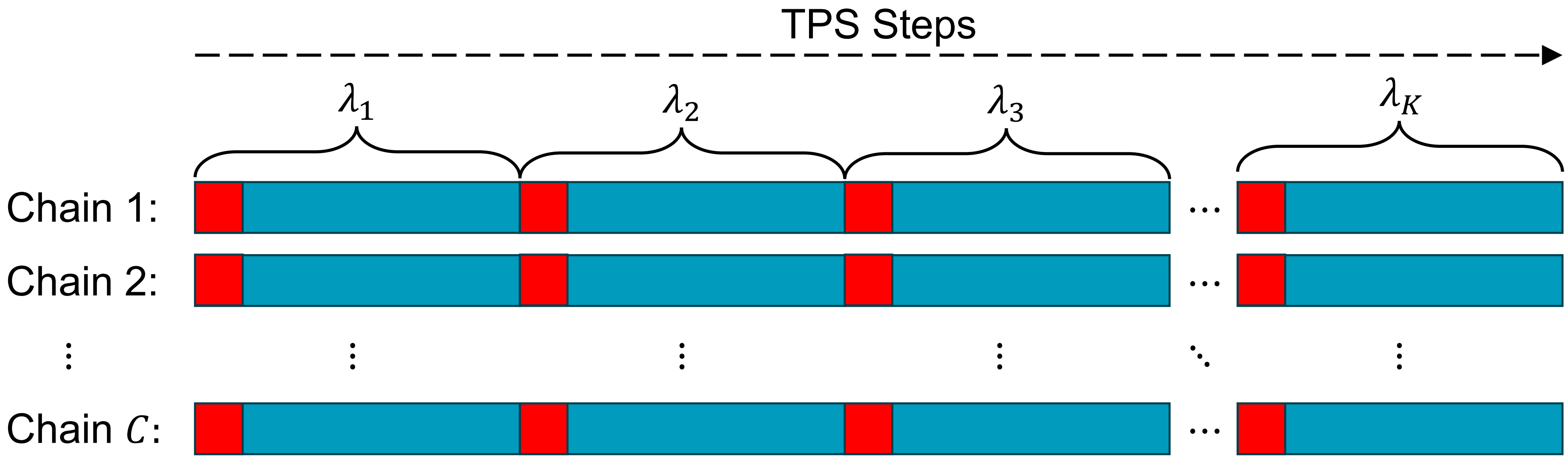
We then repeat this process with C IID chains. Hence, we generate:



Annealing & Removing Unconverged Samples

We use annealing, which means varying λ as the chain progresses.

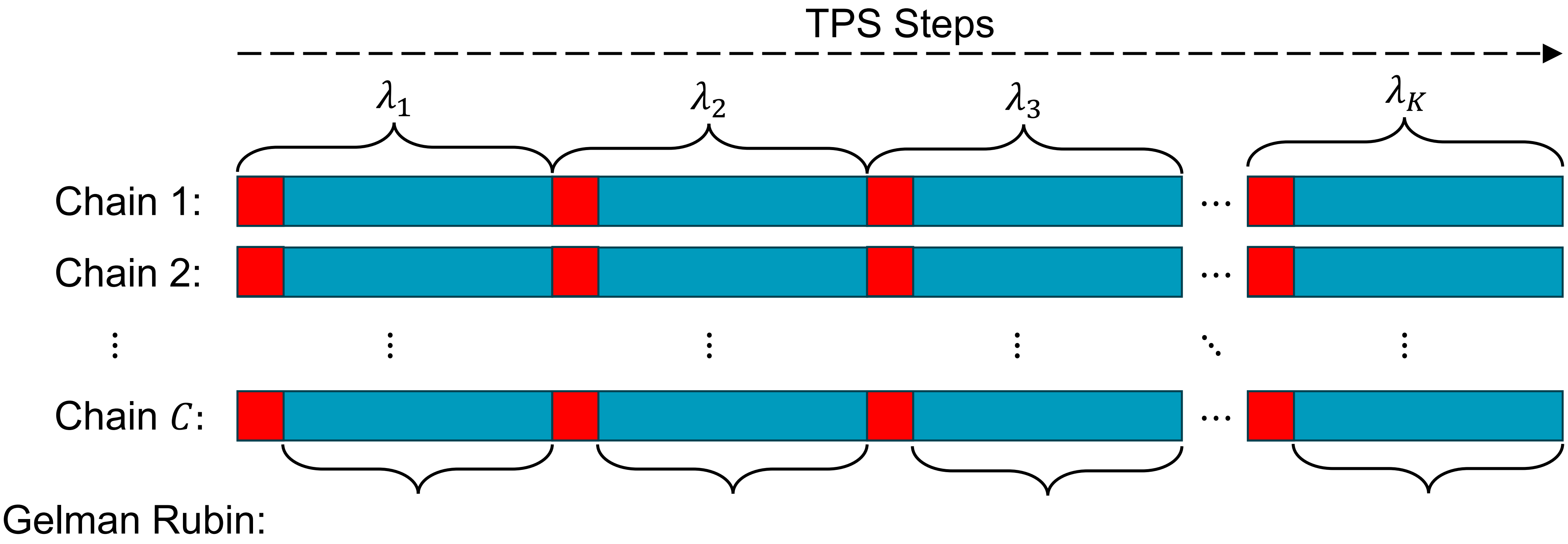
We then repeat this process with C IID chains. Hence, we generate:



Annealing & Removing Unconverged Samples

We use annealing, which means varying λ as the chain progresses.

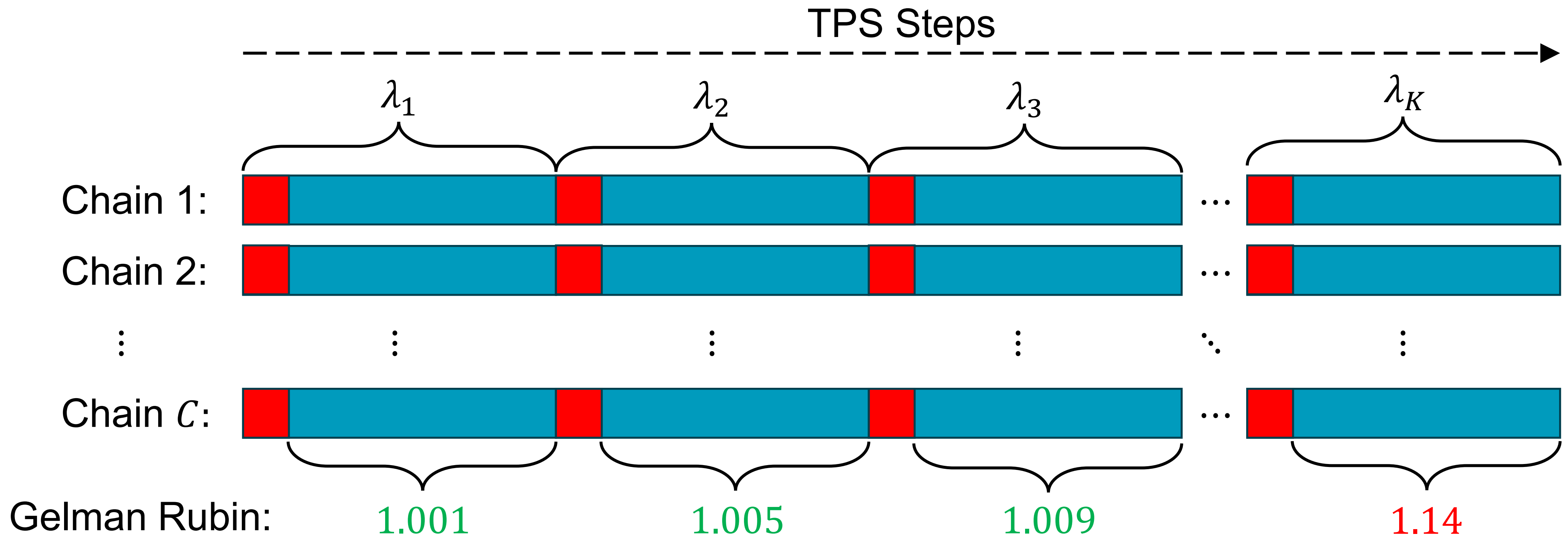
We then repeat this process with C IID chains. Hence, we generate:



Annealing & Removing Unconverged Samples

We use annealing, which means varying λ as the chain progresses.

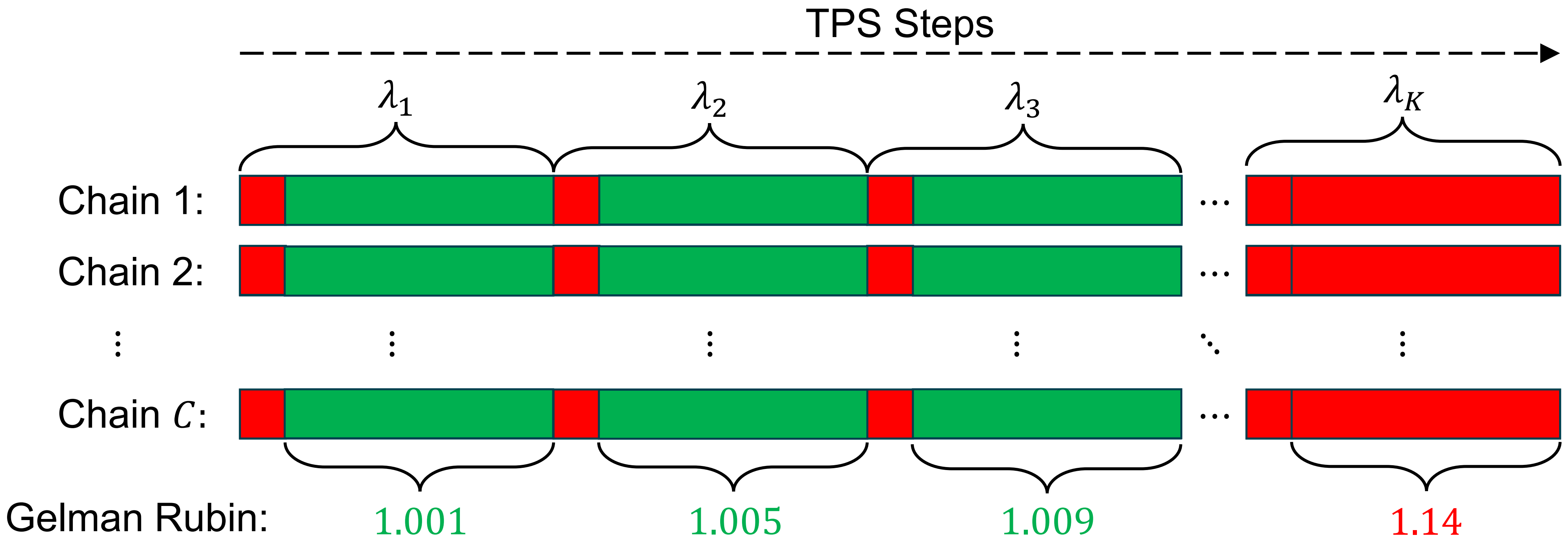
We then repeat this process with C IID chains. Hence, we generate:



Annealing & Removing Unconverged Samples

We use annealing, which means varying λ as the chain progresses.

We then repeat this process with C IID chains. Hence, we generate:



Importance Sampling the Exponential Distribution

After removing unconverged data, we are left with a matrix of data:

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1^1 & \cdots & \phi_S^1 \\ \vdots & \ddots & \vdots \\ \phi_1^K & \cdots & \phi_S^K \end{pmatrix}.$$

We then use the Multistate Bennett Acceptance Ratio (MBAR) to estimate $Z(\lambda)$:

$$\hat{Z}_1, \dots, \hat{Z}_K = \mathrm{MBAR}(\mathbf{\Phi}).$$

This is then used to calculate the importance sampling weights.

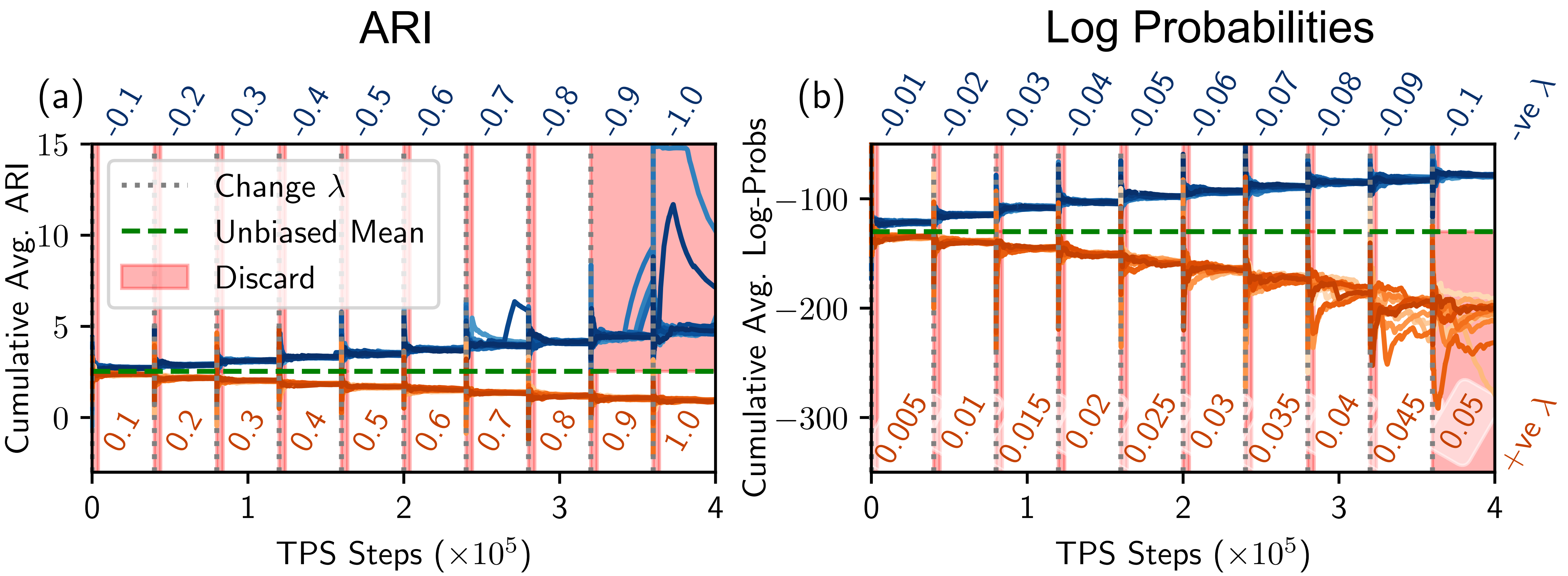
Results

In our work, we demonstrate this methodology on a TinyStories model, a Language Model trained on children's stories.

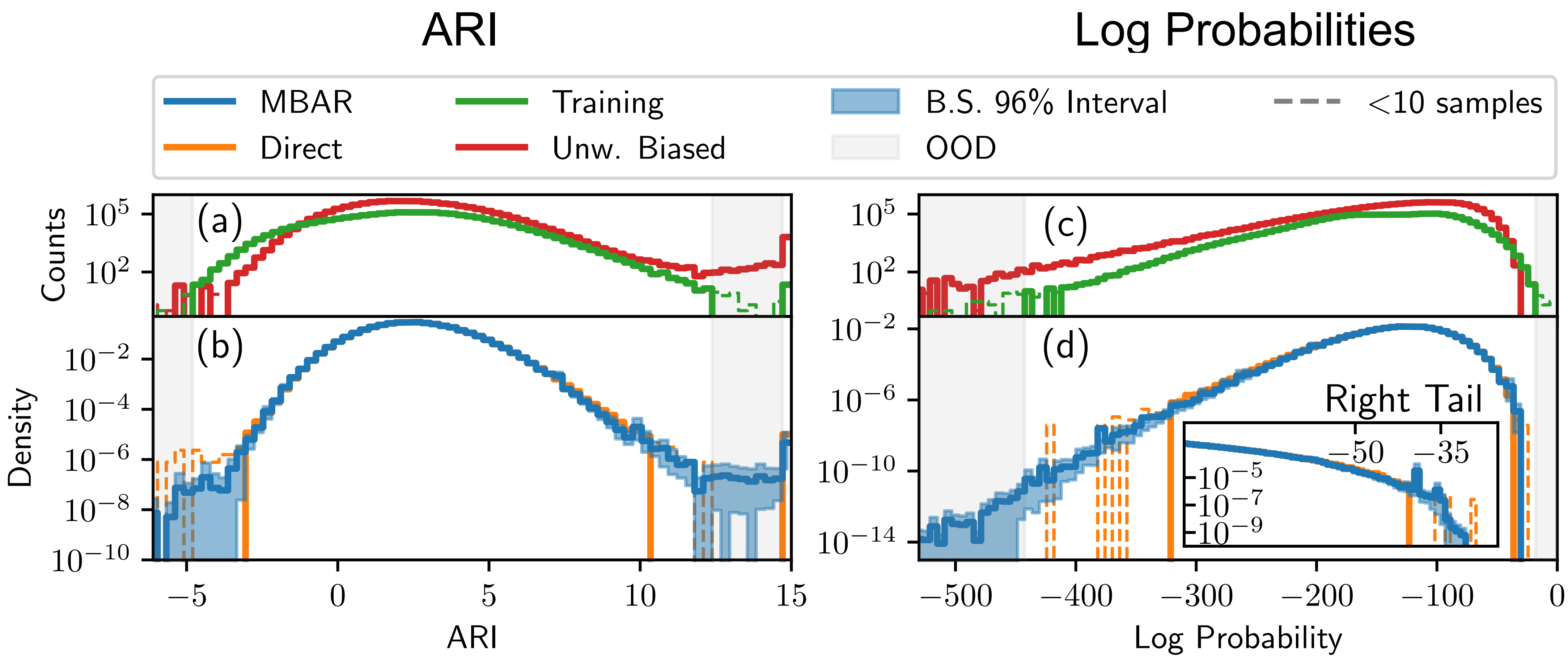
We study two metrics:

- The Automated Readability Index (ARI).
- The log probability of the completion, under the target model.

Results


ARI
Log Probabilities
(a)
(b)
-0.1
-0.2
-0.3
-0.4
-0.5
-0.6
-0.7
-0.8
-0.9
-1.0
Change λ
Unbiased Mean
Discard
Cumulative Avg. ARI
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
TPS Steps ($\times 10^5$)
-0.01
-0.02
-0.03
-0.04
-0.05
-0.06
-0.07
-0.08
-0.09
-0.1
-ve λ
Cumulative Avg. Log-Probs
0.005
0.01
0.015
0.02
0.025
0.03
0.035
0.04
0.045
0.05
+ve λ
TPS Steps ($\times 10^5$)

Results



ARI
Log Probabilities
MBAR
Direct
Training
Unw. Biased
B.S. 96% Interval
OOD
<10 samples
(a)
(b)
(c)
(d)
Counts
Density
Right Tail
ARI
Log Probability

Conclusion

- We have developed a framework to sample and estimate the likelihood of rare, unwanted completions in LLMs.
- This framework does **not require access to model weights**, allowing evaluation of closed-source models.
- We have estimated probabilities as low as 10^{-14} using only $\sim 10^6$ samples, a $\mathbf{10^8}$ **improvement over Direct Sampling.**
- In the paper, we provide rigorous error analysis of the histograms.
- We then perform EDA on the generated rare completions, to demonstrate how these methods could inform the development of filters for unwanted completions.

Future Work

Future work includes:

- Applying this to an AI Safety metric for a larger model of interest.
- Investigating alternative proposals for TPS:
  - Infilling models,
  - Finetuning to target exponential reweighting,
  - Model steering.
- Further error and asymptotic analysis of MCMC in the case of limited samples.

ICML
International Conference
On Machine Learning

Thank you!

Questions?

Poster: #1508
Hall A, Today @ 2:30 PM – 4:15 PM

Email: Jake.Dorman@nottingham.ac.uk

Further Resources