SAMPLING ALGORITHMS FOR MACHINE LEARNING WITH AUXILIARY RANDOM VARIABLES AND DIFFUSION MODELS

CURTIS MCDONALD

SEPT 12, 2025

SAMPLING PROBLEMS

- Want to produce samples from a density

$$p(w) \propto p_0(w)e^{\beta L(w)}$$

- Ex. Bayesian models/ posterior densities, numerical integration, statistical physics, generative models
- The densities can have complex structure, multi-model, non-concave, that makes them difficult to sample
- Traditional Markov Chain Monte Carlo (MCMC) is not cutting it
- New need algorithms tailored to modern problems

NEW ALGORITHMS: TWO APPROACHES

AUXILIARY RANDOM VARIABLES

DIFFUSION MODELS

AUXILIARY RANDOM VARIABLES

- Target variable w, target density $p(w)$
- Any joint density $p(w, \xi)$ with $p(w) = \int p(w, \xi) d\xi$ is fine for sampling
- ξ is ``auxiliary'' random variable, user defined only for sampling purposes.
- Can help put structure on joint density easier to sample from
- Ex. Hamiltonian MC, Simulated Annealing/ Tempering, Proximal Sampling

Gibb's Sampling

- Conditional densities $p(w|\xi), p(\xi|w)$
- Alternate sampling conditionals
- What is mixing time of this MCMC?

Mixture Representation

- $p(w) = \int p(w|\xi)p(\xi)d\xi$
- Sample $\xi \sim p(\xi), w \sim p(w|\xi)$, gives draw of $w \sim p(w)$
- Can we establish when both $p(\xi), p(w|\xi)$ are easy to sample?

- Auxiliary r.v. conditional normal, $\xi|w \sim N(\lambda\, w, I), \xi = \lambda\, w + (1-\lambda)Z, Z \sim N(0,1)$
- "Noisy" version of target random variable

LOG-CONCAVE COUPLING

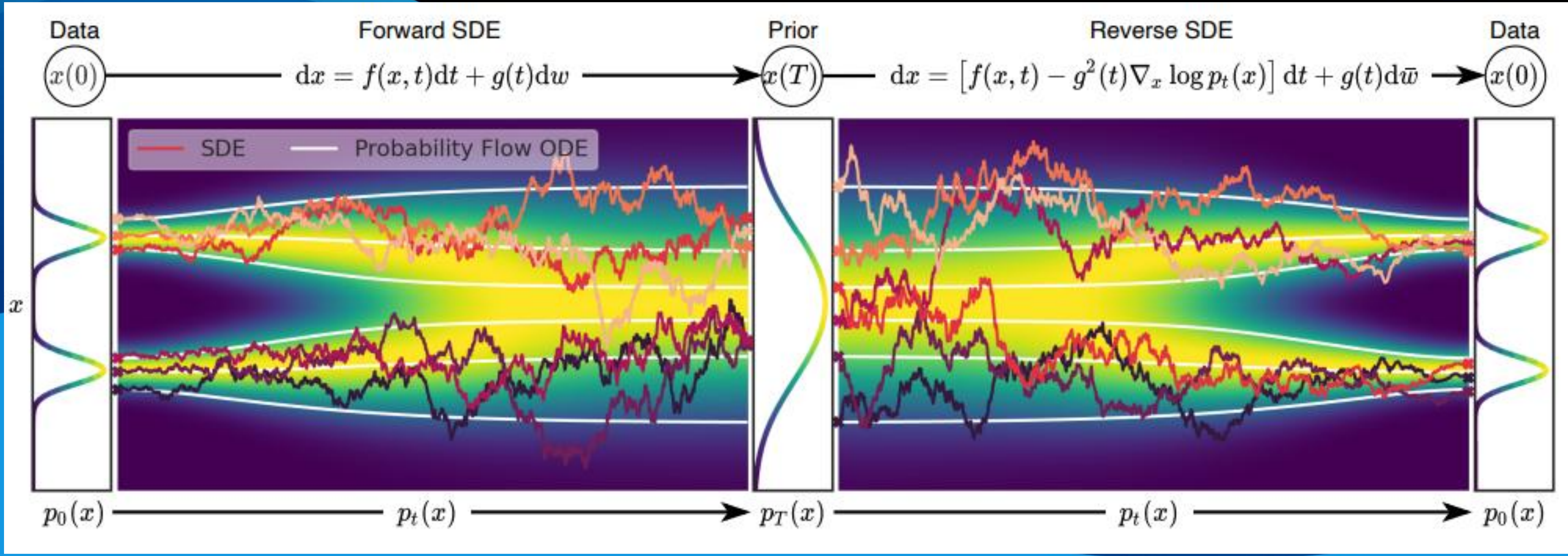
- Given target density $p(w)$, a **log-concave coupling** is a joint density $p(w, \xi)$ such that

Satisfies 3 properties	1. Target marginal is maintained, $p(w) = \int p(w\|\xi)p(\xi)d\xi$ 2. For all ξ, the reverse conditional density $p(w\|\xi)$ is log-concave 3. Auxiliary marginal density $p(\xi)$ is log-concave
Mixture Representation	Mixture density with log-concave mixing density $p(\xi)$, log-concave mixture components $p(w\|\xi)$
Easy Sampling	Easy sampling of $\xi \sim p(\xi)$ with off the shelf MCMC, followed by draw $p(w\|\xi)$ also easy

DIFFUSION MODELS

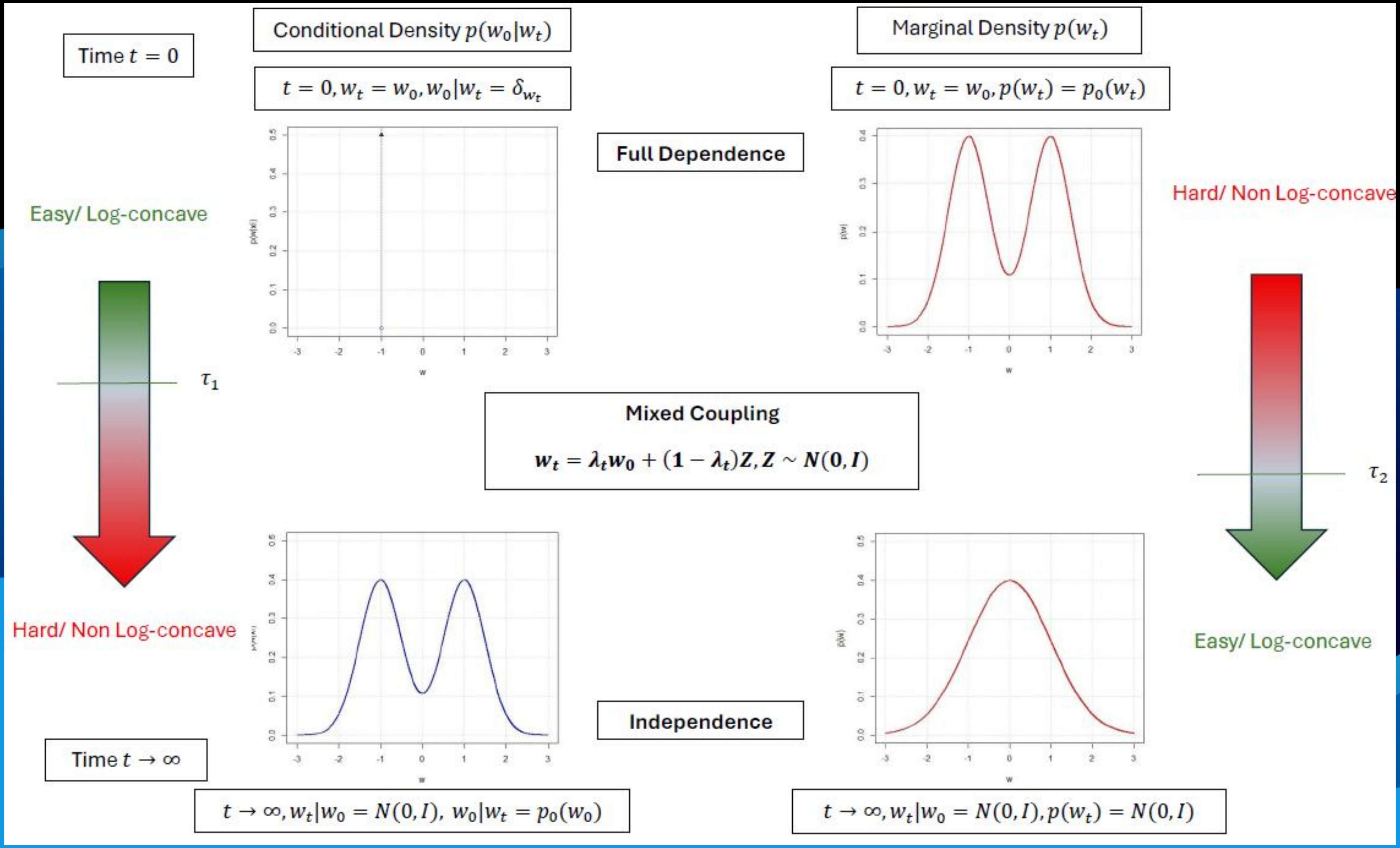
- Forward SDE moves from target to simple normal
- Reverse moves from simple to target
- Need scores of forward marginals $\nabla \log p_t(w_t)$ to implement reverse flow
- Each joint pair $p(w_0, w_t)$ is mixture representation $p^*(w_0) = \int p(w_0|w_t)p(w_t)dw_t$
- OU Process tied to normal conditionals

$$p(w_t|w_0) \sim N(e^{-t}w_0, (1-e^{-2t})I) \leftrightarrow w_t = \lambda w_0 + (1-\lambda)Z, Z \sim N(0,1)$$

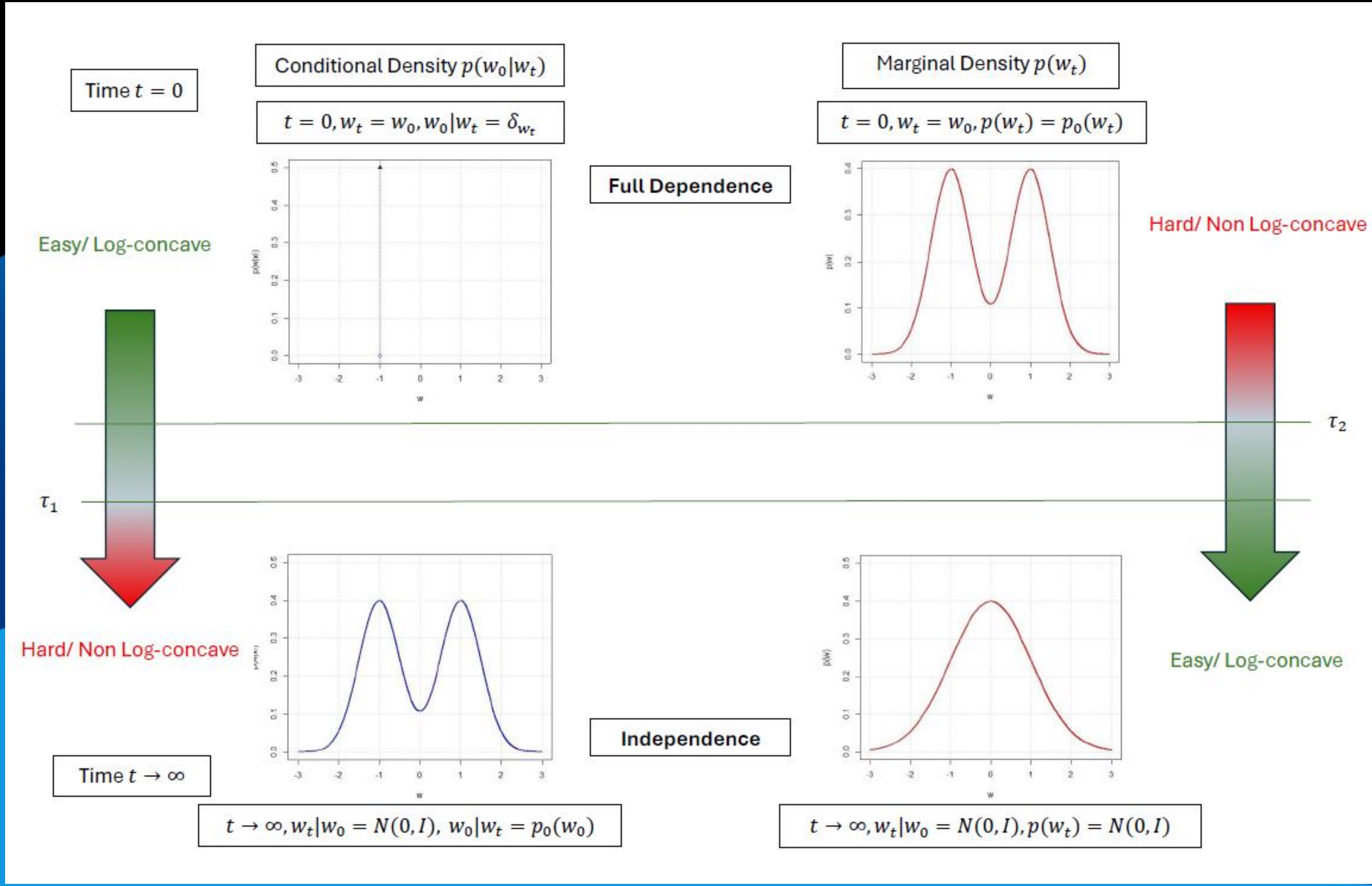


[1]

EVOLUTION OF DENSITY



EASY REGIME: LOG-CONCAVE COUPLING



- If conditional threshold happens after marginal threshold
- Exist times $\tau_2 < t < \tau_1$ such that both problems are easy at same time
- Mixture representation
  $$p(w_0) = \int p(w_0|w_t)p(w_t)dw_t$$
- Log-concave coupling
- Easily sampled by MCMC
- "One-shot" reverse diffusion

OVER PARAMETERIZED NEURAL NETWORKS

- Over parameterized neural networks have this mixture representation
- Single hidden layer neural network, K neurons each weight vector dimension d.

$$f(x,w) = V \sum_{k=1}^{K} \frac{1}{K} \varphi(x \cdot w_k)$$

- Only trian inner weights, Kd parameters overall
- N data pairs $(x_i, y_i)_{i=1}^{N}$, gain β, posterior density

$$p(w) \propto p_0(w) e^{-\beta \sum_{i=1}^{N} (y_i - f(x_i, w))^2}$$

- **Density $p(w)$ has log-concave coupling when (number of parameters) = $Kd > (\beta N)$^2**
- $p(w) = \int p(w|\xi) p(\xi) d\xi$, $p(\xi)$ log-concave and $p(w|\xi)$ log-concave

INTERPRETATION OF RESULT

- Condition: $p(w) \propto p_0(w)e^{-\beta \sum_{i=1}^{N}(y_i - f(x_i,w))^2}, Kd > (\beta N)^2$

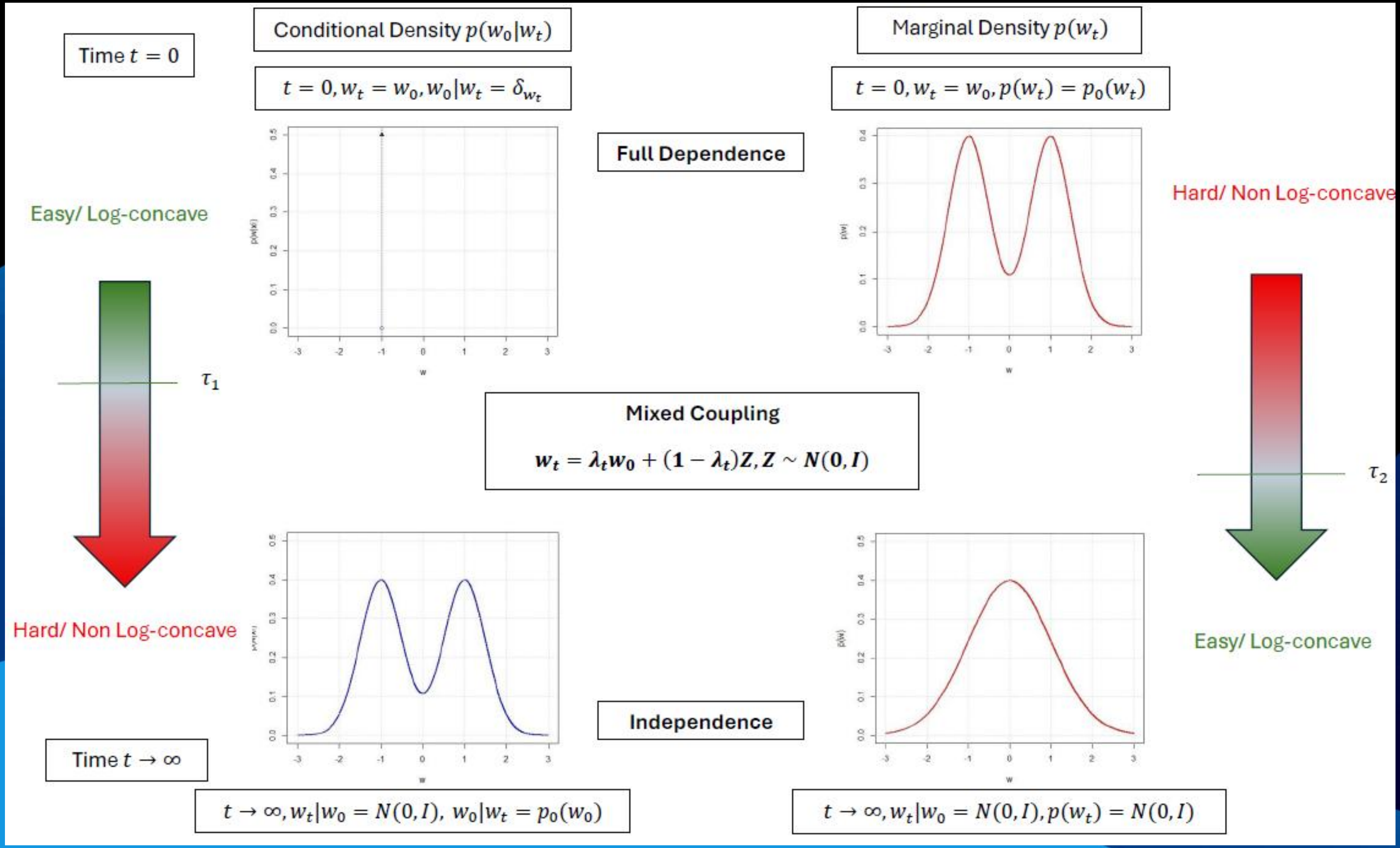
  $$p(w) \propto p_0(w)e^{-(\beta N)\sum_{i=1}^{N}\frac{1}{N}(y_i - f(x_i,w))^2}, \lambda = \beta N, \mathrm{Kd} > \lambda^2$$

- For fixed N and gain β (or fixed λ), as we increase number of parameters network will eventually enter log-concave coupling regime

- Rearrange for $\beta, \beta < \frac{\sqrt{Kd}}{N}$ or $\lambda < \sqrt{Kd}$

- β is on a spectrum from 0 to infinity. "Easy" sampling up to $\frac{\sqrt{Kd}}{N}$

HARD CASE: NO OVERLAP



- If "easy" regions don't overlap, no single simple expression for target density
- Have to run full reverse diffusion model
- Need to compute scores of forward SDE
- Ongoing research
- For NN:
  - $Kd < (\beta N)^2$ (less parameters)
  - $\beta > \sqrt{(Kd)}/N$ (high gain)

SUMMARY

- Sampling problems of interest today require new algorithms
- Auxiliary random variables can provide structure for MCMC
- Diffusion models define mixture representations of target density
- For single hidden layer NN, easy mixture when overparameterized
- Can provide insight for loss landscape and increase interest in Bayesian methods in ML

PAPER ON TOPIC

McDonald, C., Barron, A.

***Rapid Bayesian Computation and Estimation for Neural Networks via Log-Concave Coupling*.**

March, 2025. arXiv.

Scan me!

REFERENCES

- [1] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." *arXiv preprint arXiv:2011.13456* (2020).