

- Deep Generative Model.

$$\mathcal{M} = \{P_\phi : \phi \in \Xi\}$$

- prescribed probabilistic models:

- from $P_\phi(x)$ directly.

$$y = f(x)$$

- implicit probabilistic models:

- Partition the randomness and the structure into two different pbm.
- Generate latent $z_i \sim p_0(z)$ and compute $x_i = g_\phi(z)$ with some g_ϕ .
- Induces a more complex model/family of distributions $P_\phi(x)$.
- e.g. $z \sim \text{Unif}(0, 1)$, $x = F_\phi^{-1}(z)$, $x \sim \text{Exp}(\beta)$.

- DGM:
 - Sample randomness from a particularly easy Dist $z \sim N(0, I)$.
 - Use a deep NN as the structure map g_ϕ .
 - induces $P_\phi(x)$.

1. Variational Inference $q_\pi(z) = \arg \min_{q \in Q} \text{KL}(q \parallel p)$.

Variational autoencoder (VAE). Kingma & Welling 2014.

2. Generative modeling

$$z_i \sim N(0, I) \rightarrow g_\phi(z_i) \rightarrow P_\phi$$

GAN Goodfellow 2015.

• Variational Inference

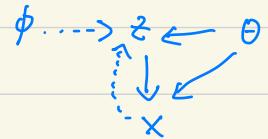
- $\mathcal{Q} = \{q_\phi(z|x) : \phi \in \Phi\}$, latent code from decoder.
- If the posterior density is $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$, then.
- $q^*(z|x) = \underset{q \in \mathcal{Q}}{\operatorname{arg\,min}} \text{KL}(q(z|x) \parallel p(z|x))$.

• ELBO objective. (evidence lower bound).

$$\begin{aligned} \text{KL}(q(z|x) \parallel p(z|x)) &= E_{q_\phi} \left(\log \frac{q(z|x)}{p(z|x)} \right) \\ &= E(\log q(z|x)) - E(\log p(z|x)) \\ &= E(\log q(z|x)) - E(\log p(z,x)) + \log p(x) \\ &\propto E_{q_\phi}(\log q_\phi(z|x)) - E_{q_\phi}(\log p(z,x)) \end{aligned}$$

$$\text{ELBO: } F(\phi, \theta) = -E_{q_\phi}(\log q_\phi(z|x)) + E_{q_\phi}(\log p_\theta(z,x)).$$

- $p_\theta(x|z)$: probabilistic decoder. $z \rightarrow x$
- $q_\phi(z|x)$: probabilistic encoder. $x \rightarrow z$.



$$p_\theta(x|z) = N(\mu_\theta(z), \sigma_\theta^2(z)), \quad \mu_\theta, \sigma_\theta^2 \text{ diag mx. are NN maps } z \rightarrow x.$$

$$q_\phi(z|x) = N(\mu_\phi(x), \sigma_\phi^2(x)) \quad \mu_\phi, \sigma_\phi^2 \text{ NN maps } x \rightarrow \mathbb{R}^d$$

- Reparameterization trick: $\varepsilon \sim N(0, \text{Id})$ and $z = \mu_\phi(x) + \sigma_\phi(x)\Theta\varepsilon \rightarrow z|x \sim N(\mu_\phi(x), \sigma_\phi^2(x))$.

calculating the expectations:

$$\arg\max_{\phi} F(\phi, \theta) = \arg\max_{\phi} \left\{ -E_{q_{\phi}}(\log q_{\phi}(z|x)) + E_{q_{\phi}}(\log P_{\theta}(z|x)) \right\}.$$

Turn to Stochastic optimization and mini-batch G.D.

- Draw a noise minibatch $\varepsilon_1, \dots, \varepsilon_M$ from $N(0, 1)$
- Draw a data minibatch x_1, \dots, x_M from dataset.
- Compute $z_m = \mu_{\phi}(x_m) + \sigma_{\phi}(x_m) \odot \varepsilon_m$
- Approximate objective.

$$\hat{F}(\phi) = \frac{1}{M} \sum_{m=1}^M \log q_{\phi}(z_m|x_m) + \frac{1}{M} \sum_{m=1}^M \log P_{\theta}(z_m, x_m).$$

- follows SGD.

Weight Uncertainty in Neural Network 2015 Google DeepMind.

- BNN. $P(y|x, w)$

for NN: MLE: $w^{\text{MLE}} = \arg \max_w \log P(D|w)$ D: data
 $= \arg \max_w \sum_i \log P(y_i|x_i, w).$

Maximum Posterior (MAP)

$$\begin{aligned} w^{\text{MAP}} &= \arg \max_w \log P(w|D) \\ &= \arg \max_w \log \frac{P(D|w) \cdot P(w)}{P(D)} \\ &= \arg \max_w \log P(D|w) + \log P(w). \end{aligned}$$

• Bayesian estimation: $P(\hat{y}|\hat{x}) = E_{p(w|D)} [P(\hat{y}|\hat{x}, w)]. = E_{q(w|\theta^*)} [P(\hat{y}|\hat{x}, w)].$

$$p(w|D) = \frac{P(D|w) \cdot P(w)}{P(D)} \quad \text{intractable.}$$

• variational inference

$$\begin{aligned} \theta^* &= \arg \min_{\theta} D_{KL}(q(w|\theta) \| p(w|D)) \\ &= \arg \min_{\theta} \int q(w|\theta) \cdot \log \frac{q(w|\theta)}{p(w) \cdot P(D|w)} dw. \end{aligned}$$

$$= \arg \min_{\theta} D_{KL}(q(w|\theta) \| p(w)) - E_{q(w|\theta)} [\log P(D|w)]$$

objective function: $F(D, \theta) = D_{KL}(q(w|\theta) \| p(w)) - E_{q(w|\theta)} [\log P(D|w)].$

Note: $-\log(\text{likelihood}) = \text{loss function.}$

$p(w)$: scale mixture gaussian prior:

$$p(w) = \prod_j \left\{ \pi \mathcal{N}(w_j | 0, \sigma_1^2) + (1-\pi) \mathcal{N}(w_j | 0, \sigma_2^2) \right\}.$$

To compute $E_{q(w|\theta)} [\log P(D|w)]$. using reparameterization.

$Z \sim N(\mu, \sigma^2)$. $Z = \sigma \varepsilon + \mu$, $\varepsilon \sim N(0, 1)$, ε , $Z \sim g_\varepsilon(\varepsilon)$. $g(\varepsilon) d\varepsilon = g(w|\theta) dw$.

$$\begin{aligned}\frac{\partial}{\partial \theta} E_{q(w|\theta)} [f(w, \theta)] &= \frac{\partial}{\partial \theta} \int f(w, \theta) \cdot q(w|\theta) dw \\ &= \frac{\partial}{\partial \theta} \int f(w, \theta) \cdot g(\varepsilon) d\varepsilon \\ &= \int \frac{\partial f(w, \theta)}{\partial \theta} \cdot g(\varepsilon) d\varepsilon. \\ &= E_{g(\varepsilon)} \left[\frac{\partial f(w, \theta)}{\partial w} \cdot \frac{\partial w}{\partial \theta} + \frac{\partial f(w, \theta)}{\partial \theta} \right].\end{aligned}$$

$$\Rightarrow \frac{\partial}{\partial \theta} E_{q(w|\theta)} [f(w, \theta)] = E_{g(\varepsilon)} \left[\frac{\partial f(w, \theta)}{\partial w} \cdot \frac{\partial w}{\partial \theta} + \frac{\partial f(w, \theta)}{\partial \theta} \right].$$

$$\text{approx: } F(D, \theta) \approx \sum_{i=1}^n \left\{ \log g(w^{(i)}|\theta) - \log P(w^{(i)}) - \log P(D|w^{(i)}) \right\}.$$

where $w^{(i)}$ denotes the Monte Carlo sample drawn from the variational posterior $g(w^{(i)}|\theta)$.

- Gaussian variational posterior.

Suppose that the variational posterior is a diagonal Gaussian Dist, a sample of weight w

$\sim \sigma Z + \mu$. $Z \sim N(0, 1)$. $\sigma = \log(1 + \exp(p))$. The variational posterior parameters are $\theta = (\mu, p)$.

$w = t(\theta, \varepsilon) = \mu + \log(1 + \exp(p)) \circ \varepsilon$. \circ : point wise multiplication.

1. Sample $\varepsilon \sim N(0, I)$.

2. let $w = \mu + \log(1 + \exp(p)) \circ \varepsilon$

3. let $\theta = (\mu, p)$.

4. let $f(w, \theta) = \log g(w|\theta) - \log P(w) \cdot P(D|w)$.

5. $\Delta \mu = \frac{\partial f(w, \theta)}{\partial w} + \frac{\partial f(w, \theta)}{\partial \mu}$

6. $\Delta p = \frac{\partial f(w, \theta)}{\partial w} \cdot \frac{\varepsilon}{1 + \exp(-p)} + \frac{\partial f(w, \theta)}{\partial p}$

7. update the variational parameters: $\mu \leftarrow \mu - \alpha \Delta \mu$; $p \leftarrow p - \alpha \Delta p$.

prior: $p(w) = \prod_j \left[\pi N(w_j | 0, \sigma_j^2) + (1-\pi) N(w_j | 0, \sigma_z^2) \right]$ w_j is the j th wgt;

- Mini-batch G.D.

$$F_i^{EQ}(D_i, \theta) = \frac{1}{M} KL[q(w|\theta) \| p(w)] - E_{q(w|\theta)} [\log P(D_i|w)].$$

$$\Rightarrow \sum_i F_i^{EQ}(D_i, \theta) = F(D, \theta).$$

or $F_i^{\pi}(D_i, \theta) = \pi_i KL[q(w|\theta) \| p(w)] - E_{q(w|\theta)} [\log P(D_i|w)].$

$\pi_i \in [0, 1]$, $\sum_{i=1}^M \pi_i = 1 \Rightarrow E[\sum_{i=1}^M F_i^{\pi}(D_i, \theta)] = F(D, \theta)$ unbiased est.

e.g. $\pi_i = \frac{2^{M-i}}{2^M - 1}$.

- local reparameterization. Variational Dropout and the Local Reparameterization Trick
2015, CIFAR.

$$Y = XW, q(w_{ij}) = N(\mu_{ij}, \sigma_{ij}^2) \quad \forall w_{ij} \in W.$$

↓

$$q(y_{m,j} | X) = N(\delta_{m,j}, \gamma_{m,j}). \quad \delta_{m,j} = \sum_i x_{m,i} \mu_{ij}, \quad \gamma_{m,j} = \sum_i x_{m,i}^2 \sigma_{ij}^2.$$

Chapter 8 Gradient flows.

- Functional $F: \mathbb{R}^d \rightarrow \mathbb{R}$, $x_0 \in \mathbb{R}^d$.

- Gradient flow Cauchy problem : $\begin{cases} x'(t) = -\nabla F(x(t)) & \text{for } t > 0, \\ x(0) = x_0. \end{cases}$

standard Cauchy pbm has a unique solution if ∇F is Lipschitz continuous, $F \in C^1$.

- Prop 8.1 : Suppose that F is convex and let x_1 and x_2 be two solution of $x'(t) = -\nabla F(x(t))$ (if F is not differentiable, we can consider $x'(t) \in \partial F(x(t))$)
Then we have $|x_1(t) - x_2(t)| \leq |x_1(0) - x_2(0)|$ for every t .
This gives uniqueness of the solution of the Cauchy pbm.

pf : let $g(t) = \frac{1}{2}|x_1(t) - x_2(t)|^2$ and differentiate it. We have

$$\frac{\partial}{\partial t} g(t) = g'(t) = (x_1(t) - x_2(t)) \cdot (x_1'(t) + x_2'(t)) = - (x_1(t) - x_2(t)) (\nabla F(x_1(t)) - \nabla F(x_2(t))).$$

By Convexity, $\forall x_1, x_2, (x_1 - x_2) \cdot (\nabla F(x_1) - \nabla F(x_2)) \geq 0$.

Generally, $\forall x_1, x_2, p_1 \in \partial F(x_1), p_2 \in \partial F(x_2), (x_1 - x_2) \cdot (p_1 - p_2) \geq 0$.

Thus, $g'(t) \leq 0$ and $g(t) \leq g(0)$. ■

• RMK 8.2: From the same proof, one can also deduce uniqueness and stability estimates in the case

where F is only semi-convex. ...

• Gradient flows discretization in time. Fix a small time step parameter $\tau > 0$ and

look for a sequence of points $(x_k^\tau)_k$ defined through.

$$x_{k+1}^\tau \in \arg \min_x F(x) + \frac{\|x - x_k^\tau\|^2}{2\tau}$$

Backward Euler: $x'(t) = -\nabla F(x(t)) \Rightarrow \frac{x_{k+1}^\tau - x_k^\tau}{\tau} + \nabla F(x_{k+1}^\tau) = 0$

$$\Rightarrow \nabla \left(\frac{\|x - x_k^\tau\|^2}{2\tau} + F(x) \right) \Big|_{x=x_{k+1}^\tau} = 0.$$

The convexity assumptions on F is not necessary for this part of the analysis.

Mild assumptions on F i.e. l.s.c. or $F(x) \geq C_1 - C_2 \|x\|^2$ lower bound. are sufficient

to guarantee that these pbm admit a solution for small τ .

• The case F is λ -convex is covered by these assumptions and also provides uniqueness of

the minimizers.

Interpret this sequence of points as the values of the curve $x(t)$ at times

$t = 0, \tau, 2\tau, \dots, k\tau, \dots$. It happens that the optimality conditions of the recursive

minimization exactly give a connection between these minimization problems and the equation.

since we have.

$$x_{k+1}^\tau \in \arg \min_x F(x) + \frac{|x - x_k^\tau|^2}{2\tau} \Rightarrow \nabla F(x_{k+1}^\tau) + \frac{x_{k+1}^\tau - x_k^\tau}{\tau} = 0.$$

$\Rightarrow \frac{x_{k+1}^\tau - x_k^\tau}{\tau} = -\nabla F(x_{k+1}^\tau)$. This expression is exactly the discrete-time

implicit Euler scheme for $\dot{x} = -\nabla F(x)$.

- ODE $\dot{x}(t) = V(x(t))$, $x(0) = x_0$. fix a time step $\tau > 0$. define a sequence x_k^τ .
 - explicit scheme : $x_{k+1}^\tau = x_k^\tau + \tau \cdot V(x_k^\tau)$, $x_0^\tau = x_0$. easy to compute, not very stable.
 - implicit scheme : $x_{k+1}^\tau = x_k^\tau + \tau \cdot V(x_{k+1}^\tau)$, $x_0^\tau = x_0$. stable, hard to compute.
- It is possible to prove that for $\tau \rightarrow 0$, the sequence we found, suitably interpolated, converges to the solution of the problem.

let (X, d) be a metric space, F is l.s.c. $F: X \rightarrow \mathbb{R} \cup \{\infty\}$ hold below.

we can define $x_{k+1}^\tau \in \arg \min_x F(x) + \frac{d(x, x_k^\tau)^2}{2\tau}$. and study limit $\tau \rightarrow 0$.

We can consider two different interpolations of the points x_k^z , given, respectively, by.

$$x^z(t) = x_k^z \quad , \quad \tilde{x}^z(t) = w_{x_{k-1}} x_k^z \left(\frac{t - (k-1)z}{z} \right) \quad \text{for } t \in [z(k-1), z, kz].$$

where $w_{x,y}(s)$ denotes any constant speed geodesic connecting a point x to a point y ,

parametrized on unit interval $[0,1]$. The interpolation \tilde{x}^z only make sense in spaces

where geodesics exists.

• Def 8.3: a curve $x: [0, T] \rightarrow X$ is said to be a minimizing movement if \exists a sequence

of time steps $\tau_j \rightarrow 0$ s.t. the piecewise constant interpolations $x^{z,j}$, built from a

sequence of solutions of the iterated minimization scheme &l. uniformly converge to x on $[0, T]$.

...

Can define a scheme on metric space (X, d) , $X = P(\Omega)$, $d = W_2$.

Ω is compact, F is l.s.c. and bdd below.

Define $x_T^{n+1} \in \arg\min \{F(x) + \frac{d(x, x^n)^2}{2\tau}\}$.

$x_T(t) = x_T^n$, study $t \rightarrow 0$.

By continuity equation $\rho_t + \nabla \cdot (\rho v) = 0$. to flow densities.

Goal: to find velocity field v s.t. this flow agrees with $\lim_{t \rightarrow 0} x_T(t)$.

* optimality condition in JKO scheme. Compute the 1st variation.

want to perturb $\rho \in P(\Omega)$ to $\rho + \varepsilon \chi$, $\rho + \varepsilon \chi \in P(\Omega)$. so $F(\rho + \varepsilon \chi)$ is well defined.

for χ , let $\sigma = \rho + \chi \in P(\Omega)$ so $\forall \varepsilon > 0$ small ε , $\rho + \varepsilon \chi = \rho + \varepsilon(\sigma - \rho) = (1-\varepsilon)\rho + \varepsilon\sigma \in P(\Omega)$.
as long as $\rho, \sigma \in P(\Omega)$.

The first variation of F , $\frac{\partial}{\partial \rho} F(\rho)$ is.

$$\frac{d}{d\varepsilon} F(\rho + \varepsilon \chi) \Big|_{\varepsilon=0} = \int \frac{\delta}{\delta \rho} F(\rho) \chi(x) dx \quad \text{for all } \chi = \sigma - \rho, \sigma \in P(\Omega) \cap L_c^{\infty}(\Omega).$$

$$\text{Note that } \int \left(\frac{\partial F}{\partial \rho} + c \right) \chi(x) dx = \int \frac{\partial F}{\partial \rho} \chi(x) dx + \underbrace{c \int \chi''(x) dx}_{\rightarrow 0}.$$

The 1st variation is defined uniquely only up to additive constants.

$$\text{let } G(p) = F(p) + \frac{W_2(p, p_z^n)^2}{2\varepsilon}$$

$$\text{need to compute } \frac{\delta}{\delta p} G(p) = \frac{\delta}{\delta p} F(p) + \frac{1}{2\varepsilon} \cdot \frac{\delta}{\delta p} W_2(p, p_z^n)^2$$

$$\text{The dual formulation: } W_2(f, g)^2 = 2 \inf_{\pi \in \Pi(f, g)} \int \frac{|x-y|^2}{2} d\pi(x, y)$$

$$= 2 \cdot \max_{u, v} \left\{ \int u f dx + \int v g dy \mid u(x) + v(y) \leq \frac{1}{2} |x-y|^2 \right\}.$$

$$= 2 \cdot \max_u \left\{ \int u f dx + \int u^c g dy \right\}.$$

$$\begin{aligned} \frac{d}{d\varepsilon} W_2(f + \varepsilon x, g)^2 \Big|_{\varepsilon=0} &= 2 \cdot \frac{d}{d\varepsilon} \max_u \left\{ \int u (f + \varepsilon x) dx + \int u^c g dy \right\} \Big|_{\varepsilon=0} \\ &= 2 \cdot \max_u \left\{ \int u x dx \right\} = 2 \cdot \int u^* x dx \end{aligned}$$

where u^* achieves the max. u^* is a Kantorovich potential for the transport with cost $\frac{1}{2}|x-y|^2$

from p to p_z^n .

When using OT, the optimal map is

$$T(x) = x - \nabla u^*(x) = x - (\nabla h)^{-1}(\nabla u^*(x)), \quad h(z) = \frac{1}{2}|z|^2.$$

$$\text{and } \frac{d}{d\varepsilon} W_2(f + \varepsilon x, g) \Big|_{\varepsilon=0} = 2 \cdot \int u^* x dx \Rightarrow \frac{\delta W_2(p, p_z^n)^2}{\delta p} = 2 u^*.$$

$T(x) = x - \nabla u^*(x)$ is the optimal map from p to p_z^n
 $\Rightarrow \nabla u^* = x - T(x)$

The JKO scheme is $\beta_T^{n+1} = \arg \min_p \{ F(p) + \frac{W_2(p, p_T^n)^2}{2\tau} \}$.

$$= \arg \min_p G(p).$$

$$\Rightarrow \frac{\delta G(\beta_T^{n+1})}{\delta p} + C = 0.$$

$$\Rightarrow \frac{\delta}{\delta p} F(\beta_T^{n+1}) + \frac{u^*}{\tau} = \text{Constant}.$$

$$\Rightarrow \nabla \left(\frac{\delta F}{\delta p} \right) + \frac{\nabla u^*}{\tau} = 0.$$

$$\Rightarrow \nabla \left(\frac{\delta F}{\delta p} \right) + \frac{x - T(x)}{\tau} = 0.$$

$$\Rightarrow \underbrace{\frac{T(x) - x}{\tau}}_{\text{Velocity.}} = V(x) = \nabla \left(\frac{\delta F}{\delta p} \right)$$

PDE associated with Gradient in W_2 .

$$\beta_t + \nabla \cdot (\beta v) = 0 \Rightarrow \beta_t + \nabla \cdot (\beta \cdot \nabla \left(\frac{\delta F}{\delta p} \right)) = 0.$$

want to perturb $p \in P(\Omega)$ to $p + \varepsilon x$, $p + \varepsilon x \in P(\Omega)$. So $F(p + \varepsilon x)$ is well defined.
 for x , let $\sigma = p + x \in P(\Omega)$ so $\forall \varepsilon > 0$ small ε , $p + \varepsilon x = p + \varepsilon(\sigma - p) = (1 - \varepsilon)p + \varepsilon\sigma \in P(\Omega)$.
 as long as $p, \sigma \in P(\Omega)$.

- e.g. $F(p) = \int p \cdot \log p \, dx$ (negative entropy) we want a flow that maximizes entropy

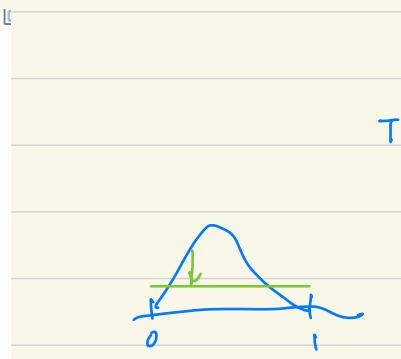
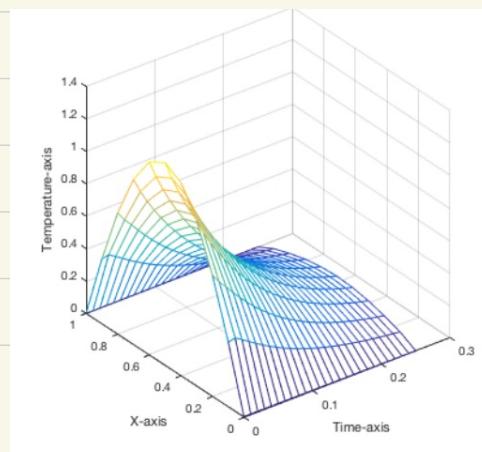
$$\begin{aligned} \frac{d}{d\varepsilon} F(p + \varepsilon x) \Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon} \int (p + \varepsilon x) \log(p + \varepsilon x) \, dx \Big|_{\varepsilon=0} \\ &= \frac{d}{d\varepsilon} \int p \log(p + \varepsilon x) + x \log(p + \varepsilon x) \, dx \Big|_{\varepsilon=0} \\ &= \int \frac{p \cdot x}{p + \varepsilon x} + x \log(p + \varepsilon x) + \frac{\varepsilon x^2}{p + \varepsilon x} \, dx \Big|_{\varepsilon=0} \\ &= \int (x \log(p) + x) \, dx. \\ \Rightarrow \frac{\delta F}{\delta p} &= \log p + 1. \end{aligned}$$

$$\nabla \left(\frac{\delta F}{\delta p} \right) = \nabla (\log p + 1) = \frac{1}{p} \cdot \nabla p.$$

Thus, the Gradient flow is: $o = p_t - \nabla (\log p + 1) = \frac{1}{p} \cdot \nabla p$
 $o = p_t - \nabla (\nabla p)$
 $p_t = \Delta p$. (heat equation).

for $p_t = \Delta p$ and $p(x, 0) = \phi(x)$. initial Dist.

$$p(x, t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} e^{-(x-y)^2/(4t)} \cdot \phi(y) \, dy. \text{ Dist at time } t.$$



$$F(q) = D_{KL}(q \parallel \pi_n) = \int \log\left(\frac{q}{\pi_n}\right) q dx = E_q[\log\left(\frac{q}{\pi_n}\right)].$$

$$\begin{aligned} \frac{d}{d\varepsilon} F(q + \varepsilon x) \Big|_{\varepsilon=0} &= \frac{d}{d\varepsilon} \int (q + \varepsilon x) \cdot \log\left(\frac{q + \varepsilon x}{\pi_n}\right) dx \Big|_{\varepsilon=0} \\ &= \frac{d}{d\varepsilon} \int q \cdot \log\left(\frac{q + \varepsilon x}{\pi_n}\right) + x \cdot \log\left(\frac{q + \varepsilon x}{\pi_n}\right) dx \Big|_{\varepsilon=0} \\ &= \int \frac{q \cdot x / \pi_n}{q + \varepsilon x} \cdot \pi_n + x \cdot \log\left(\frac{q + \varepsilon x}{\pi_n}\right) + \frac{\varepsilon x^2 / \pi_n}{q + \varepsilon x} \cdot \pi_n dx \Big|_{\varepsilon=0} \\ &= \int x + x \log\left(\frac{q}{\pi_n}\right) dx \end{aligned}$$

$$\Rightarrow \frac{\delta F}{\delta q} = \log\left(\frac{q}{\pi_n}\right) + 1.$$

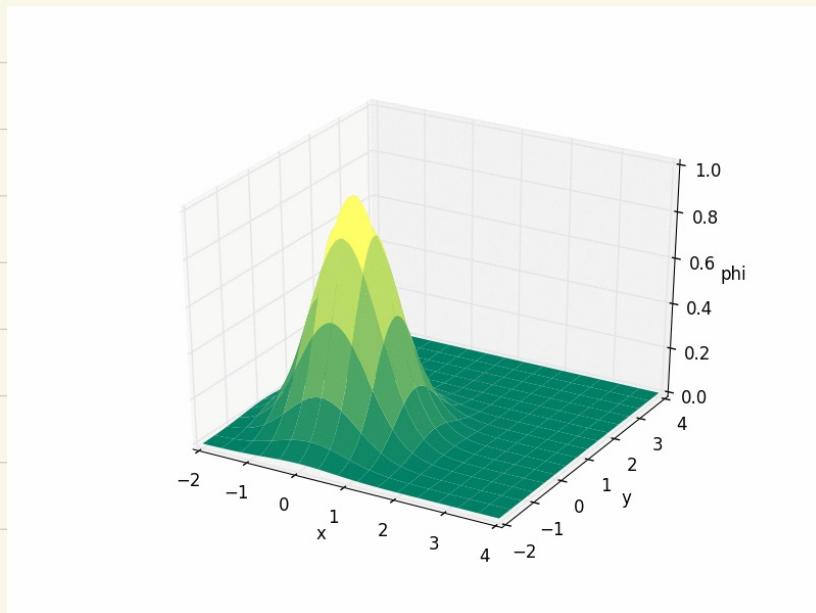
$$\nabla\left(\frac{\delta F}{\delta q}\right) = \nabla\left(\log\left(\frac{q}{\pi_n}\right) + 1\right) = \frac{\pi_n}{q} \cdot \nabla q \cdot \nabla \pi_n$$

Thus, the Gradient flow is: $0 = q_t - \nabla\left(q \cdot \frac{\pi_n}{q} \cdot \nabla q \cdot \nabla \pi_n\right)$

$$0 = q_t - \nabla \pi_n \cdot \Delta q \cdot \Delta \pi_n$$

$$q_t = \nabla \pi_n \cdot \Delta q \cdot \Delta \pi_n.$$

Advection. $q_t = \nabla \cdot (q \cdot \nabla V)$ where $V = \log \frac{q}{\pi_n}$.



$$\arg \min_{\theta} \quad p = p(w|\theta)$$

$$\arg \min_p D_{KL}(p \parallel \pi_n), \quad \pi_n = P(w|D). \quad w: \text{weights}, \quad D: \text{dataset} \quad D = (X, Y).$$

M: source Dist.

$$= \arg \min_p D_{KL}(p \parallel P(w|D))$$

$$= \arg \min_p \int p(x) \cdot \log \left[\frac{p(x)}{P(w) \cdot P(D|w)} \cdot P(D) \right] dx$$

$$\pi_n(\theta) = P(\theta|D)$$

$$= \arg \min_p \int p(x) \cdot \log \left[\frac{p(x)}{P(w) \cdot P(D|w)} \right] dx$$

$$= \arg \min_p \int p(x) \cdot \log \left[\frac{p(x)}{P(w)} \right] dx - \int \log P(D|w) \cdot p(x) dx$$

$$= \arg \min_p D_{KL}(p \parallel P(w)) - E_p [\log P(D|w)].$$

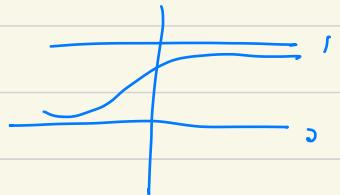
$$\text{or } = \arg \min_p \int p(x) \cdot \log p(x) dx + \underbrace{\int \log(P(D|w) \cdot P(w)) \cdot p(x) dx}_{V} \dots \text{Fokker-Planck}$$

log is concave $\Rightarrow V$ is convex.

Define $F(p) = (*)$.

$$p(y|x, \beta) =$$

$$\partial_t p = \Delta p + \nabla \cdot (p \cdot \nabla V).$$



$$(p)^T_{t+1} \in \arg \min_{p \in \mathcal{W}_2(\mathcal{K})} F(p) + \frac{1}{2\tau} W_2^2(p, p_t^T).$$

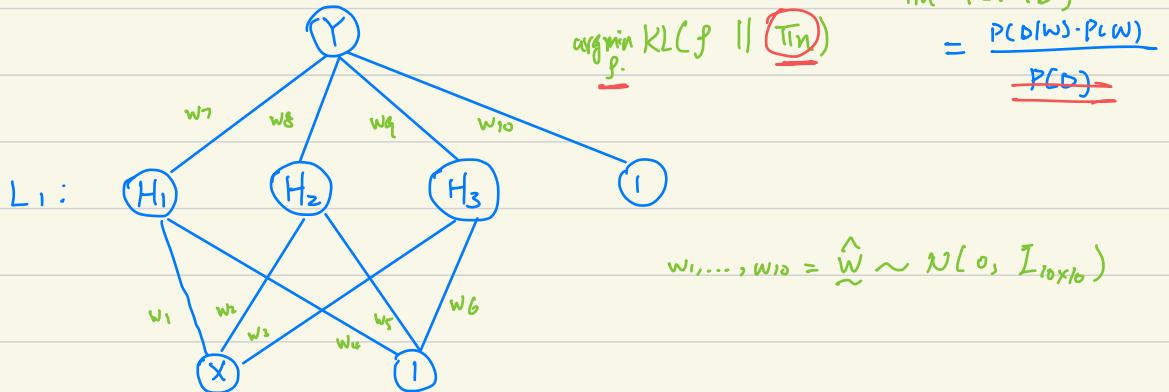
$$\text{prob}(y=1|x) = \text{logit}(x^T \beta).$$

$$x \in \mathbb{R}^{n \times d}, \beta \in \mathbb{R}^d$$

$$P(\beta|x, y), \quad P(\beta) \sim N(0, \sigma^2 I_{d \times d}).$$

$$P(y|x, \beta) = \prod_{i=1}^n P(y_i|x_i, \beta).$$

$$x \in \mathbb{R}, \quad y \in \mathbb{R}. \quad Y = f(x) + \varepsilon.$$



H_1, H_2, H_3 activative function $\sigma(\cdot)$. $D = (\underline{x}, \underline{y})$.

Given x_i . $L_1: H_1 = \sigma(\hat{w}_1 x_i + \hat{w}_4); H_2 = \sigma(\hat{w}_2 x_i + \hat{w}_5); H_3 = \sigma(\hat{w}_3 x_i + \hat{w}_6)$

$$Y: Y = H_1 \cdot w_7 + H_2 \cdot w_8 + H_3 \cdot w_9 + w_{10}. \Rightarrow P(\hat{Y}_i | x_i, \hat{w})$$

$$P(\hat{Y} | \underline{x}, \hat{w}) = \prod_{i=1}^n P(\hat{Y}_i | x_i, \hat{w})$$

$$? P(\hat{Y} | \underline{x}, \hat{w})$$

True y from data.

$$P(\hat{w} | \underline{x}, \underline{y}) = \frac{P(\underline{y} | \underline{x}, \hat{w}) \cdot P(\hat{w})}{\text{constant}}$$

$$y = x \cdot \beta + \varepsilon.$$

$$P(\beta | D) = \frac{P(D|\beta) \cdot P(\beta)}{P(D)}.$$