- Deep Generative Model.

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

  - 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 $q_*(z) = \arg\min_{q \in Q} KL(q \,\|\, p)$.
   Variational autoencoder (VAE). Kingma & Welling 2014.

2. Generative modeling

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

   GAN Goodfellow 2015.

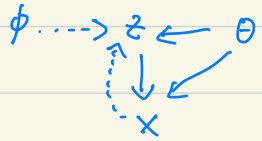
- Variational Inference
  - $Q = \{ q_\phi(z|x) : \phi \in \Phi \}$.
  - If the posterior density is $p(z|x) = \frac{p(x|z)p(z)}{p(x)}$ (latent code from data), then.

$$q^*(z|x) = \arg\min_{q \in Q} KL(q(z|x) \,||\, p(z|x)).$$

- ELBO objective. (evidence lower bound).

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

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



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

$p_\theta(x|z) = N(\mu_\theta(z), \sigma_\theta^2(z))$. $\mu_\theta, \sigma_\theta^2$ diag mx. are NN maps $z \to x$.

$q_\phi(z|x) = N(\mu_\phi(x), \sigma_\phi^2(x))$ μ_ϕ, σ_ϕ^2 NN maps $x \longrightarrow \mathbb{R}^d$

- Reparameterization trick: $\varepsilon \sim N(0, Id)$ and $z = \mu_\phi(x) + \sigma_x(x) \odot \varepsilon \longrightarrow z|x \sim N(\mu_\phi(x), \sigma_\phi^2(x))$.

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 $\epsilon_1, \ldots, \epsilon_M$ from $N(0, I)$
- Draw a data minibatch $x_1, \ldots, x_M$ from dataset.
- Compute $z_m = \mu_\phi(x_m) + \sigma_\phi(x_m) \odot \epsilon_m$
- 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: $w^{MLE} = \arg\max_w \log P(D|w)$ D: data

$$= \arg\max_w \sum_i \log P(y_i|x_i,w).$$

Maximum Posteriori (MAP)

$$w^{MAP} = \arg\max_w \log P(w|D)$$

$$= \arg\max_w \log \frac{P(D|w)\cdot P(w)}{P(D)}$$

$$= \arg\max_w \log P(D|w) + \log P(w).$$

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

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

- variational inference

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

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

complexity cost — likelihood cost

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

Note: $-\log$(likelihood) = 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 , $z \sim q_z(\varepsilon)$. $q(\varepsilon)d\varepsilon = q(w|\theta)dw$.

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

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

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

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

- Guassian variational posterior.

suppose that the variational posterior is a diagonal Guassian Dist, a sample of weight $w \sim \sigma\cdot Z + \mu$. $Z \sim N(0,1)$. $\sigma = \log(1+\exp(\rho))$. The variational posterior parameters are $\theta = (\mu, \rho)$. $w = t(\theta, \varepsilon) = \mu + \log(1+\exp(\rho)) \circ \varepsilon$. $\circ$: point wise multiplication.

1. Sample $\varepsilon \sim N(0, I)$.
2. let $w = \mu + \log(1+\exp(\rho)) \circ \varepsilon$
3. let $\theta = (\mu, \rho)$.
4. let $f(w,\theta) = \log q(w|\theta) - \log P(w)\cdot P(D|w)$.
5. $\Delta_\mu = \frac{\partial f(w,\theta)}{\partial w} + \frac{\partial f(w,\theta)}{\partial \mu}$
6. $\Delta_\rho = \frac{\partial f(w,\theta)}{\partial w}\cdot\frac{\varepsilon}{1+\exp(-\rho)} + \frac{\partial f(w,\theta)}{\partial \rho}$
7. update the variational parameters: $\mu \leftarrow \mu - \alpha\Delta_\mu$; $\rho \leftarrow \rho - \alpha\Delta_\rho$.

prior: $P(\underset{\sim}{w}) = \prod_j [\pi N(w_j|0,\sigma_1^2) + (1-\pi)N(w_j|0,\sigma_2^2)]$ w_j is the j^{th} wgt ;

- Mini-batch G.D.

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

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

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

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

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

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

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

$$\Downarrow$$

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

Chapter 8 Gradient flows.

- Funtional $F : \mathbb{R}^d \to \mathbb{R}$, $x_0 \in \mathbb{R}^d$.

- 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, $F \in C^{1,1}$.

- 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 $|x_1(t) - x_2(t)| \le |x_1(0) - x_2(0)|$ 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, $\forall x_1, x_2$, $(x_1 - x_2) \cdot (\nabla F(x_1) - \nabla F(x_2)) \ge 0$.

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

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

• 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 r.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 pbm and the equation.

since we have.

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

$$\Rightarrow \frac{x^\tau_{k+1} - x^\tau_k}{\tau} = -\nabla F(x^\tau_{k+1}).$$ This expression is exactly the discrete-time implicit Euler scheme for $x' = -\nabla F(x)$.

- ODE $x'(t) = V(x(t))$. $x(0) = x_0$. fix a time step $\tau > 0$. define a sequence x^τ_k.
  - explicit scheme : $x^\tau_{k+1} = x^\tau_k + \tau \cdot V(x^\tau_k)$, $x^\tau_0 = x_0$. easy to compute, not very stable.
  - implicit scheme : $x^\tau_{k+1} = x^\tau_k + \tau \cdot V(x^\tau_{k+1})$, $x^\tau_0 = x_0$. stable, hard to compute.

- It is possible to prove that for $\tau \to 0$, 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. $F : X \to \mathbb{R} \cup \{+\infty\}$ bdd below.

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

we can consider two different interpolations of the points x_k^τ, given, respectively, by.

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

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 $\tilde{x}^\tau$ only make sense in spaces where geodesics exists.

- Def 8.3: a curve $x : [0,T] \to X$ is said to be a minizing movement if $\exists$ a sequence of time steps $\tau_j \to 0$ s.t. the piecewise constant interpolations x^{τ_j}, built from a sequence of solutions of the iterated minization scheme 8.1. 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 $X_\tau^{n+1} \in \arg\min \left\{ F(x) + \frac{d(x, x^n)^2}{2\tau} \right\}$.

$X_\tau(t) = X_\tau^n$, study $\tau \to 0$.

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 $\lim_{\tau \to 0} X_\tau(t)$.

• 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 define.

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, $\frac{\delta}{\delta \rho} F(\rho)$ is.

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

Note that $\int \left(\frac{\delta F}{\delta \rho} + c\right) \chi(x)\, dx = \int \frac{\delta F}{\delta \rho} \chi(x)\, dx + c \underbrace{\int \chi(x)\, dx}_{\substack{\sigma - \rho \\ \| \\ \to 0.}}$

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

let $G(\rho) = F(\rho) + \frac{W_2(\rho, \rho_\tau^n)^2}{2\tau}$

need to compute $\frac{\delta}{\delta \rho} G(\rho) = \frac{\delta}{\delta \rho} F(\rho) + \frac{1}{2\tau} \cdot \frac{\delta}{\delta \rho} W_2(\rho, \rho_\tau^n)^2$

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

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

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

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

$$= 2 \max_u \left\{ \int u \chi dx \right\} = 2 \cdot \int u^* \chi dx$$

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

When using OT, the optimal map is

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

and $\frac{d}{d\varepsilon} W_2(f+\varepsilon\chi, g)\Big|_{\varepsilon=0} = 2 \cdot \int u^* \chi dx \Rightarrow \frac{\delta W_2(\rho, \rho_\tau^n)^2}{\delta \rho} = 2u^*$.

$T(x) = x - \nabla u^*(x)$ is the optimal map from ρ to $\hat{\rho}_\tau^n$.

$\Rightarrow \nabla u^* = x - T(x)$

The JKO scheme is $\rho_\tau^{n+1} = \arg\min_\rho \left\{ F(\rho) + \frac{W_2(\rho, \rho_\tau^n)^2}{2\tau} \right\}$.

$$= \arg\min_\rho G(\rho).$$

$$\Rightarrow \frac{\delta G(\rho_\tau^{n+1})}{\delta \rho} + C = 0.$$

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

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

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

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

works out as $\tau \to 0$.

PDE associated with Gradient in W_2.

$$\rho_t + \nabla \cdot (\rho v) = 0 \Rightarrow \rho_t + \nabla \cdot \left(\rho \cdot \nabla\left(\frac{\delta F}{\delta \rho}\right)\right) = 0.$$

want to perturb $\rho \in P(\Omega)$ to $\rho + \varepsilon\chi$, $\rho + \varepsilon\chi \in P(\Omega)$. So $F(\rho + \varepsilon\chi)$ is well define.
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 $\rho, \sigma \in P(\Omega)$.

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

$$\frac{d}{d\varepsilon} F(\rho + \varepsilon\chi)\Big|_{\varepsilon=0} = \frac{d}{d\varepsilon} \int (\rho + \varepsilon\chi) \log(\rho + \varepsilon\chi)\, dx \Big|_{\varepsilon=0}$$

$$= \frac{d}{d\varepsilon} \int \rho \log(\rho + \varepsilon\chi) + \varepsilon\chi \log(\rho + \varepsilon\chi)\, dx \Big|_{\varepsilon=0}.$$

$$= \int \frac{\rho \cdot \chi}{\rho + \varepsilon\chi} + \chi \log(\rho + \varepsilon\chi) + \frac{\varepsilon\chi^2}{\rho + \varepsilon\chi}\, dx \Big|_{\varepsilon=0}.$$

$$= \int (\chi \log(\rho) + \chi)\, dx.$$

$$\Rightarrow \frac{\delta F}{\delta \rho} = \log \rho + 1.$$

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

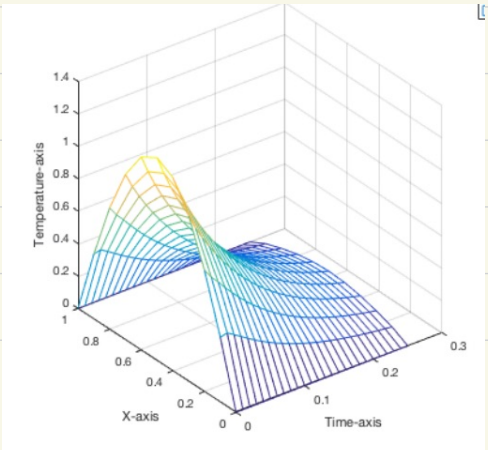
Thus, the Gradient flow is: $0 = \rho_t - \nabla(\rho \cdot \frac{1}{\rho} \cdot \nabla\rho)$

$$0 = \rho_t - \nabla(\nabla\rho)$$

$$\rho_t = \Delta\rho. \quad \text{(heat equation)}.$$

for $\rho_t = \Delta\rho$ and $\rho(x, 0) = \phi(x)$. initial Dist.

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



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

$$\frac{d}{d\varepsilon} F(q+\varepsilon x)\Big|_{\varepsilon=0} = \frac{d}{d\varepsilon}\int (q+\varepsilon x)\cdot \log\left(\frac{q+\varepsilon x}{\pi_n}\right) dx\Big|_{\varepsilon=0}$$

$$= \frac{d}{d\varepsilon}\int q\cdot\log\left(\frac{q+\varepsilon x}{\pi_n}\right) + \varepsilon x\cdot\log\left(\frac{q+\varepsilon x}{\pi_n}\right) dx\Big|_{\varepsilon=0}.$$

$$= \int \frac{q\cdot x/\pi_n}{q+\varepsilon x}\cdot\pi_n + x\cdot\log\left(\frac{q+\varepsilon x}{\pi_n}\right) + \frac{\varepsilon x^2/\pi_n}{q+\varepsilon x}\cdot\pi_n \, dx\Big|_{\varepsilon=0}$$

$$= \int x + x\log\left(\frac{q}{\pi_n}\right) dx$$

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

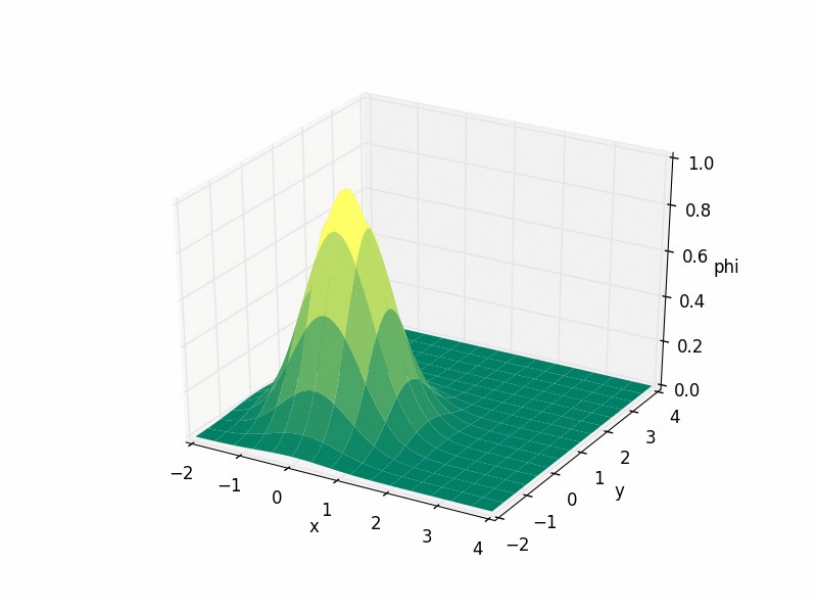
$$\nabla\left(\frac{\delta F}{\delta q}\right) = \nabla\left(\log\left(\frac{q}{\pi_n}\right) + 1\right) = \frac{\pi_n}{q}\cdot\nabla q : \nabla\pi_n$$

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

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

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

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



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

$$\arg\min_p D_{KL}(p \,\|\, \pi_n), \qquad \pi_n = P(w|D).$$ w: weights, D: dataset $D = (\underset{\sim}{X}, \underset{\sim}{Y})$. μ: source Dist.

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

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

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

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

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

$$= \arg\min_p D_{KL}(p \,\|\, P(w)) - \mathbb{E}_p[\log P(D|w)].$$

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

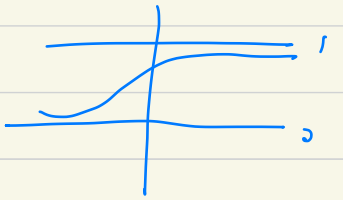
$\log$ is concave $\Rightarrow V$ is convex.

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

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

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

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



$prob(y=1|x) = logit(x^T\cdot\beta).$

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

$P(\beta|x,y)$, $P(\beta) \sim N(0, \sigma^2 I_{d\times d})$ $\sigma^2 = 10, 100$.

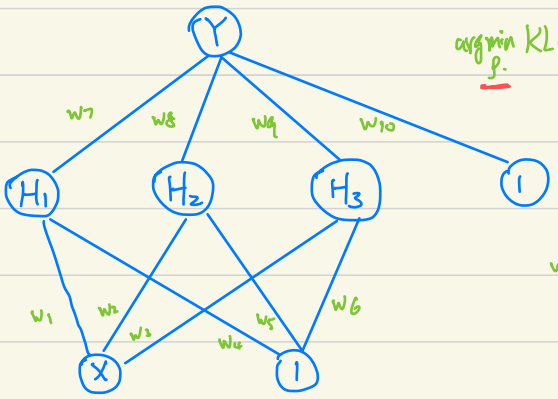
$$P(y|x,\beta) = \prod_{i=1}^n P(y_i|x_i,\beta).$$

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

$\pi_n = P(W|D) = \frac{P(D|W) \cdot P(W)}{P(D)}$

$\arg\min_f KL(f \,||\, \pi_n)$

L_1:



$w_1, \ldots, w_{10} = \hat{\underset{\sim}{w}} \sim N(0, I_{10\times 10})$

H_1, H_2, H_3 activative function $\sigma(\cdot)$. $D = (\underset{\sim}{X}, \underset{\sim}{Y})$.

Given x_1. L_1: $H_1 = \sigma(\hat{w}_1 x_1 + \hat{w}_4)$; $H_2 = \sigma(\hat{w}_2 x_1 + \hat{w}_5)$; $H_3 = \sigma(\hat{w}_3 x_1 + \hat{w}_6)$

Y: $Y = H_1 \cdot w_7 + H_2 \cdot w_8 + H_3 \cdot w_9 + w_{10}$. $\Rightarrow P(\hat{Y}_i | x_1, \hat{\underset{\sim}{w}})$

$P(\hat{\underset{\sim}{Y}} | \underset{\sim}{X}, \hat{\underset{\sim}{w}}) = \prod_{i=1}^{n} P(\hat{Y}_i | x_i, \hat{\underset{\sim}{w}})$? $P(\underset{\sim}{Y} | \underset{\sim}{X}, \underset{\sim}{w})$

$\underset{\sim}{Y}$: True y from data.

$P(\hat{\underset{\sim}{w}} | \hat{\underset{\sim}{Y}}, \underset{\sim}{X}) = \frac{P(\hat{\underset{\sim}{Y}} | \underset{\sim}{X}, \hat{\underset{\sim}{w}}) \cdot P(\hat{\underset{\sim}{w}})}{\text{constant}}$

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

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

$p(x_i | x_{i-1}) = \frac{P(x_{i-1} | x_i) \cdot P(x_i)}{P(x_{i-1})}$

$\Rightarrow P(x_{i-1} | x_i) = \frac{P(x_{i-1})}{P(x_i)} \cdot P(x_i | x_{i-1})$.