All You Need Is Relative Information

Shaowei Lin (stealth startup)

Math Machine Learning Seminar
MPI MIS + UCLA
20210909

How do we train recurrent networks?

Static Systems

Maximum Likelihood

Truth	$q(x)dx$
Data	$D_n = \{X_1, \ldots, X_n\}$
Model	$p(x\|w)dx$
Prior	$\varphi(w)dw$

Maximize generalized log-likelihood

$$\sum_{i=1}^{n} \log p(X_i|w) + a_n \log \varphi(w)$$

Maximum Likelihood

Minimize generalized log-likelihood ratio

$$R_n(w) = \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)} - a_n \log \varphi(w)$$

Maximum likelihood estimate

$$\hat{w} = \min_{w \in W} R_n(w)$$

Estimated density

$$p^*(X) = p(X|\hat{w})$$

Relative Information

(KL divergence, relative entropy)

$$K(w) \coloneqq I_{q \,\|\, p(\cdot|w)}(X) = \int q(x) \log \frac{q(x)}{p(x|w)}\, dx$$

Log-likelihood ratio, normalized training error

$$K_n(w) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{q(X_i)}{p(X_i|w)}$$

$$\mathbb{E}[K_n(w)] = K(w)$$

$$\frac{1}{n} R_n(w) = K_n(w) + \frac{a_n}{n} \log \varphi(w)$$

Generalization Error

Normalized test error

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{q(X_i^*)}{p^*(X_i^*)}$$

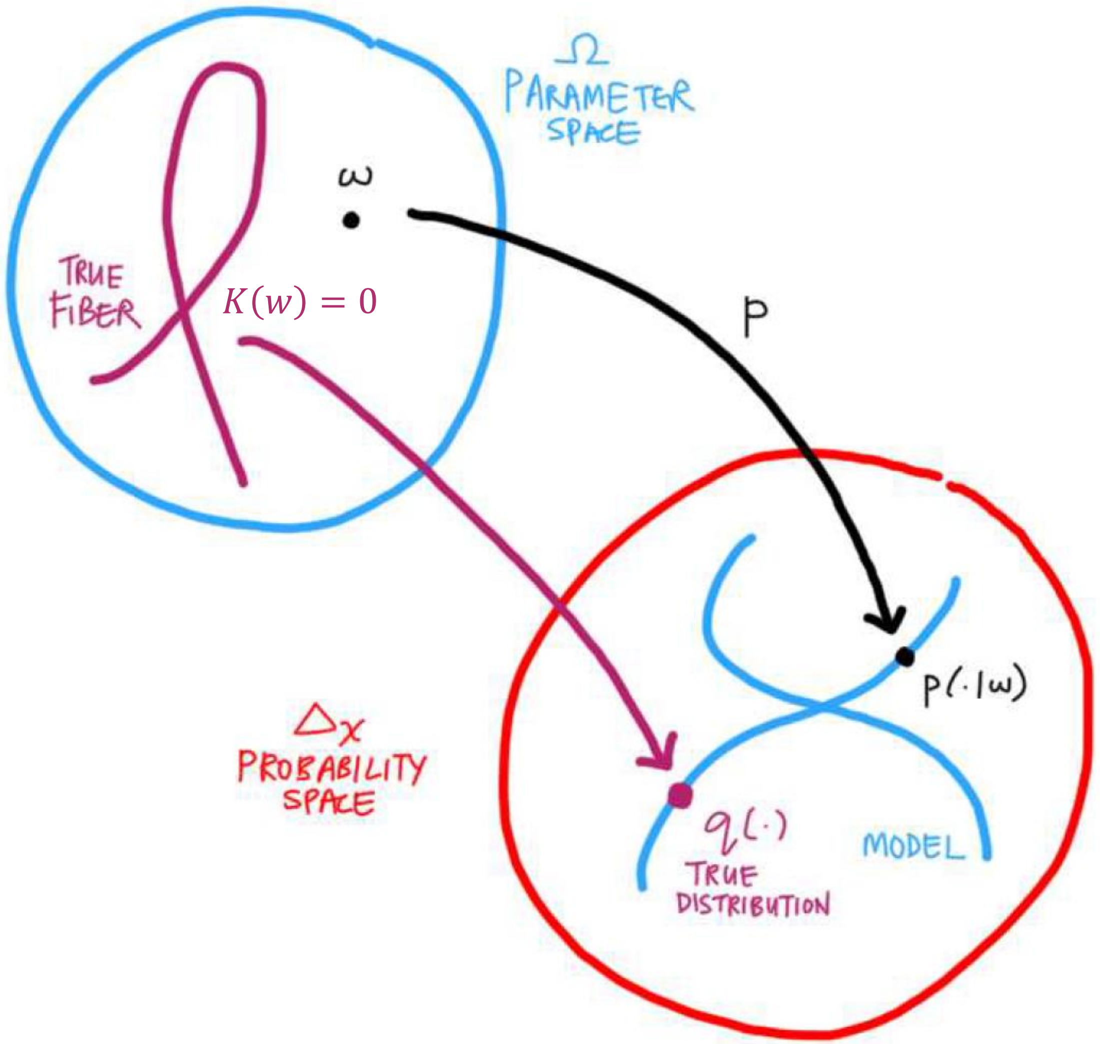
Generalization error of estimated density $p^*(x)$

$$I_{q \,\|\, p^*}(X) = \int q(x) \log \frac{q(x)}{p^*(x)}\, dx$$

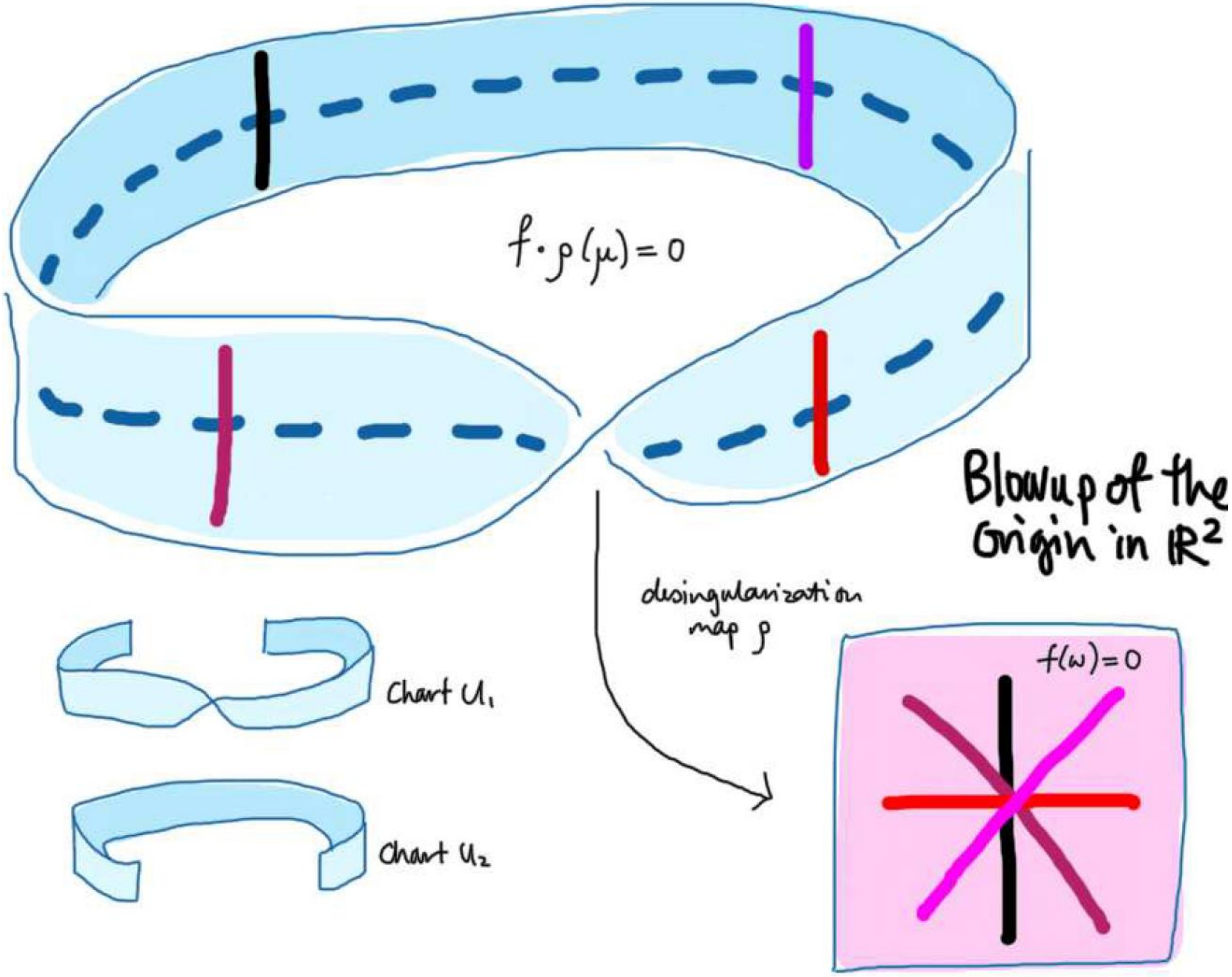
Generalization error for maximum likelihood

$$R_g = I_{q \,\|\, p(\cdot|\widehat{w})}(X) = K(\widehat{w})$$

Geometry of Singular Models



Resolution of Singularities



Singularities of $K(w)$

Resolution of singularities $\rho : M \to W$

Locally (in each chart of ρ)

$$K \circ \rho(\mu) = \mu^{2k} := \mu_1^{2k_1} \mu_2^{2k_2} \cdots \mu_d^{2k_d}$$

Standard form of log-likelihood ratio (Watanabe)

$$K_n \circ \rho(\mu) = \mu^{2k} - \frac{1}{\sqrt{n}} \mu^k \xi(\mu) + o_p\left(\frac{1}{n}\right)$$

- Gaussian process $\xi(\mu)$ on manifold M
- Random variable $o_p\left(\frac{1}{n}\right)$ with $n\, o_p\left(\frac{1}{n}\right) \to 0$ in probability

Watanabe, Sumio. *Algebraic geometry and statistical learning theory*. No. 25. Cambridge university press, 2009.

Asymptotic Generalization Error

Apply resolution of singularities $\rho: \mu \mapsto w$ and new local coordinates $(t, v_1, \dots, v_{d-1}) \mapsto (\mu_1, \dots, \mu_d)$ where $t = \mu^{2k}$

Generalized log-likelihood ratio

$$\frac{1}{n} R_n(t, v) = t^2 - \frac{1}{\sqrt{n}} t\, \xi(0, v) + \frac{a_n}{n} \log \varphi(0, v) + o_p\left(\frac{1}{n}\right)$$

Asymptotic generalization error (for $a_n = 0$)

$$\mathbb{E}[R_g] = \frac{1}{4n} \mathbb{E}\left[\max_{\mu:\, K(\mu)=0} \max\{0, \xi(\mu)\}^2\right] + o\left(\frac{1}{n}\right)$$

Watanabe, Sumio. *Algebraic geometry and statistical learning theory*. No. 25. Cambridge university press, 2009.

Bayesian Inference

Posterior distribution

$$p(w|D_n) = \frac{p(w)p(D_n|w)}{p(D_n)} = \frac{p(w)\frac{p(D_n|w)}{q(D_n)}}{\frac{p(D_n)}{q(D_n)}} = \frac{1}{Z_n^0}\varphi(w)e^{-nK_n(w)}$$

Normalized marginal likelihood $Z_n^0 = \int \varphi(w)e^{-nK_n(w)}dw$

Estimated density $p^*(X) = \int p(X|w)p(w|D_n)\,dw$

$$= \frac{\int p(X|w)p(D_n|w)p(w)dw}{\int p(D_n|w)p(w)dw}$$

$$= \frac{\int p(X,D_n|w)p(w)dw}{\int p(D_n|w)p(w)dw}$$

Bayesian Inference

Generalization error

$$B_g = \int q(x) \log \frac{q(x)}{p^*(x)} dx$$

$$= \int q(x) \log \frac{\int \frac{p(D_n|w)}{q(D_n)} p(w) dw}{\int \frac{p(x,D_n|w)}{q(x,D_n)} p(w) dw} dx$$

$$= \int q(x) \log \frac{Z_n^0}{Z_{n+1}^0} dx$$

$$= \log Z_n^0 - \mathbb{E}_{X_{n+1}}\left[\log Z_{n+1}^0\right]$$

Expected generalization error

$$\mathbb{E}\left[B_g\right] = \mathbb{E}[\log Z_n^0] - \mathbb{E}\left[\log Z_{n+1}^0\right]$$

Zeta Function

Laplace integral $Z(n) = \int \varphi(w) e^{-nK(w)} dw$

Zeta function $\zeta(z) = \int \varphi(w) K(w)^{-z} dw$

Example. If $\varphi(w) = 1$, $K \circ \rho(\mu) = \mu^{2k}$, $\rho'(\mu) = \mu^h$, then locally in the chart $[0,1]^d$

$$\zeta(z) = \int_{[0,\,1]^d} \mu^{-2kz+h} d\mu = \frac{1}{(-2k_1 z + h_1 + 1) \cdots (-2k_d z + h_d + 1)}$$

Poles are of the form $\lambda_i = \frac{h_i + 1}{2k_i}$ possibly with multiplicity

Real Log Canonical Threshold

Real log canonical threshold of $K(w)$ consists of the smallest pole λ of $\zeta(z)$ and its multiplicity m

Convergence of stochastic complexity (Watanabe)

$$\log Z_n^0 = -\lambda \log n + (m-1) \log \log n + F^R(\xi) + o_p(1)$$

Generalization error of Bayesian inference

$$\mathbb{E}[B_g] = \mathbb{E}[\log Z_n^0] - \mathbb{E}[\log Z_{n+1}^0] \approx \frac{\lambda}{n} + o\left(\frac{1}{n}\right)$$

Conjecture. For singular models, $\mathbb{E}[R_g] \gg \mathbb{E}[B_g]$

Watanabe, Sumio. *Algebraic geometry and statistical learning theory*. No. 25. Cambridge university press, 2009.

Flatness of Minima

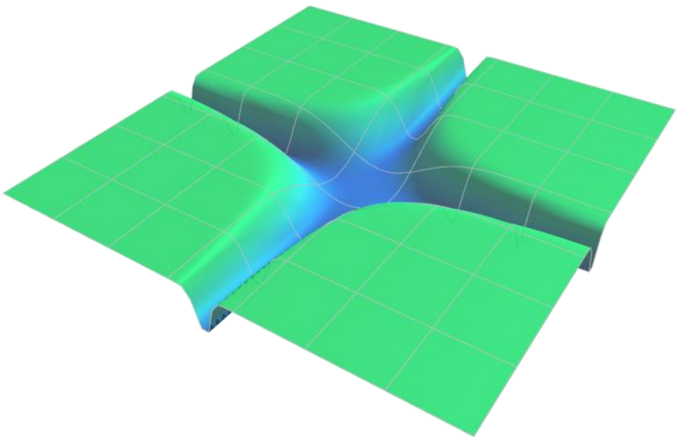
Volume of tubular neighborhood

$$V(n) = \int_{K(w)<\frac{1}{n}} \varphi(w)dw$$

$$\log V(n) = -\lambda \log n + (m-1)\log\log n + C + o(1)$$

Model selection criteria

- Smallest generalization error
- Smallest real log canonical threshold
- Largest $K(w)$-neighborhood
- Flatness/curvature not good enough



D. Murfet, S. Wei, M. Gong, H. Li, J. Gell-Redman, and T. Quella. "Deep Learning is Singular, and That's Good." *arXiv:2010.11560* (2020).

Variational Inference

Chain Rule

Conditional relative information

$$I_{q\|p}(Z|X) \coloneqq \int q(x) \int q(z|x) \log \frac{q(z|x)}{p(z|x)} dz dx$$

Chain rule $I_{q\|p}(Z,X) = I_{q\|p}(Z|X) + I_{q\|p}(X)$

Corollary $I_{q\|p}(Z,X) \geq I_{q\|p}(X)$

Variational Inference

Goal. Minimize $I_{q\|p}(X)$ over $p(X)$

Strategy. Minimize upper bound $I_{q\|p}(Z,X)$

1. Fix p and optimize over discriminative $q(Z|X)$
2. Fix q and optimize over generative $p(Z,X)$, often approximately by sampling x from $q(X)$

Example. Expectation-maximization

1. Optimal $q(Z|X)$ is $p(Z|X)$
2. E-step: $L(p|x) = \int q(z|x) \log p(z,x)\, dz$
   M-step: Maximize $L(p|x)$ over $p(Z,X)$

Maximum Likelihood

Goal. Minimize $I_{q\|p}(X)$

1. True density $q(x)$ is fixed so nothing to do
2. Find w that minimizes $K(w) = I_{q\|p(\cdot|w)}(X)$

Maximum likelihood method
Sample $K_n(w)$, compute $\nabla K_n(w)$ and descend.

Stochastic approximation method
Compute $\nabla K(w)$, sample $[\nabla K]_n(w)$ and descend.
Tends to explore w with large $K(w)$-neighborhoods.

Bayesian Inference

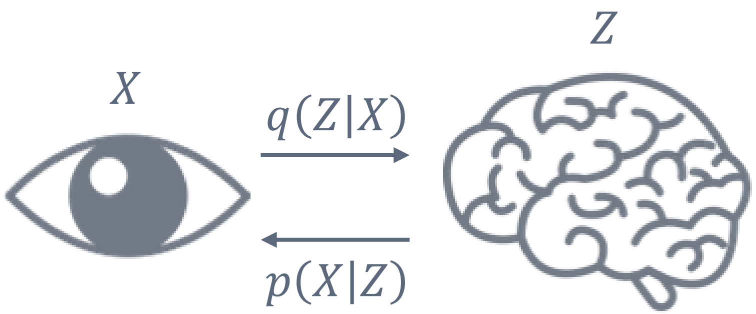
Goal. Minimize $I_{q\|p}(w, X)$

1. Optimal $q(w|X)$ is posterior $p(w|X)$
2. Find $\hat{p}(w)$ that minimizes
$$I_{q\|p}(w, X) = \int q(w|x)q(x) \log \frac{q(w|x)q(x)}{\hat{p}(w)p(x|w)} dwdx$$

Sampling x from $q(X)$, last step reduces to maximizing
$$\int q(w|x) \log \hat{p}(w)\, dw$$
where the optimal $\hat{p}(w)$ is $q(w|x) = p(w|x)$.

Hence, generative prior is updated with posterior.

Compute Perspective



- Distribution $q(X)$ of *sensor* X is *immutable*. Distribution $q(Z|X)$ of *memory* Z is *mutable*.

- Conditionals $q(Z|X)$, $p(X|Z)$ as (stochastic) *computations*. Discriminative $q(Z|X)$ infers structures from observations. Generative $p(X|Z)$ predicts observations from structures.

- $I_{q||p}(Z|X) = I_{q||p}(Z,X) - \ I_{q||p}(X)$
  Cost of structural learning completely determined by ability to invert generative $p(X|Z)$ and compute $p(Z|X)$.

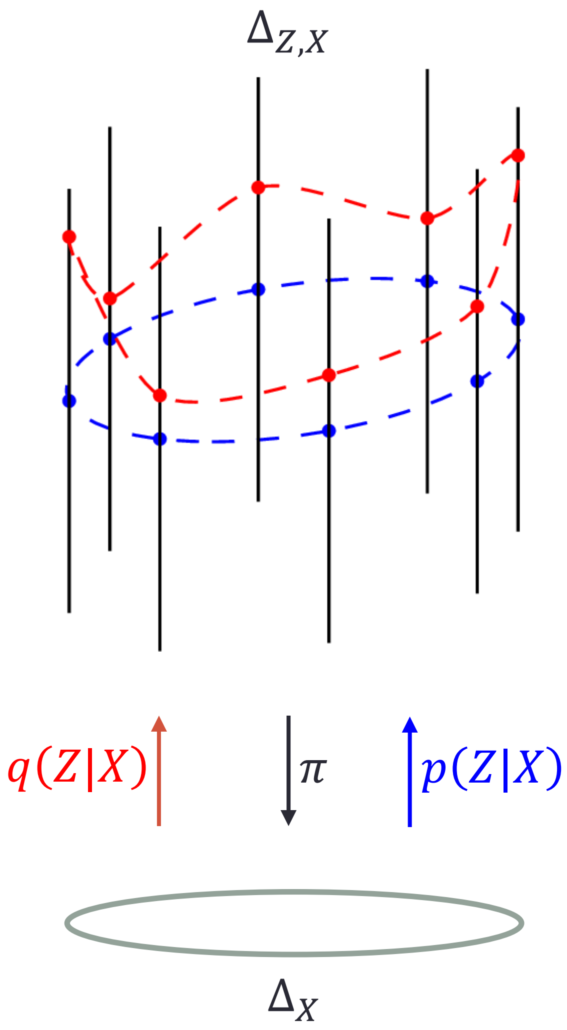
Sheaf Perspective

Space Δ_Y of distributions over Y

Bundle $\pi : \Delta_{Z,X} \to \Delta_X$
by marginalization

Sections $p(Z|X)\colon \Delta_X \to \Delta_{Z,X}$
by multiplication

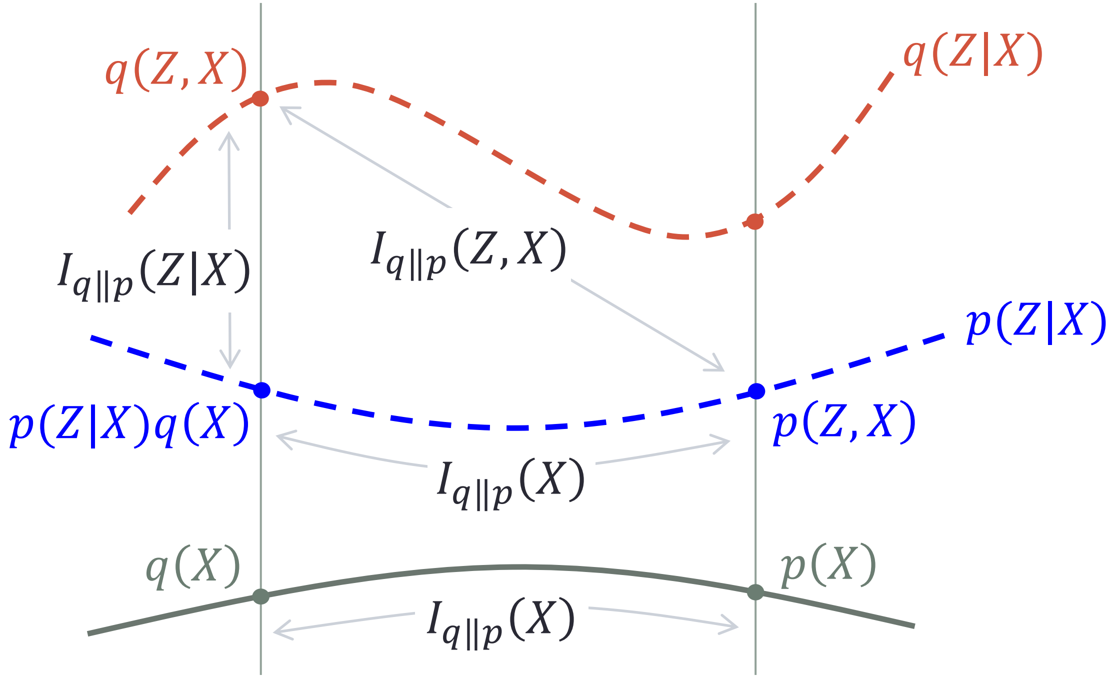
Lift optimization
of $I_{q\|p}(X)$ over Δ_X
to $I_{q\|p}(Z,X)$ over $\Delta_{Z,X}$



Sheaf Perspective

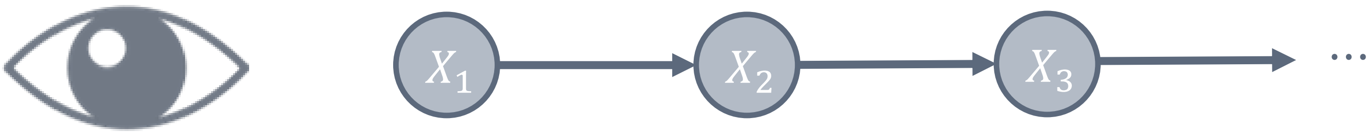
$I_{q\|p}(Z,X)$	distance to point $q(Z,X)$ from point $p(Z,X)$
$I_{q\|p}(Z\|X)$	distance to point $q(Z,X)$ from section $p(Z\|X)$
$I_{q\|p}(X)$	distance to point $q(X)$ from point $p(X)$

$$I_{q\|p}(Z,X) = I_{q\|p}(Z|X) + I_{q\|p}(X)$$



Dynamic Systems

Stochastic Processes



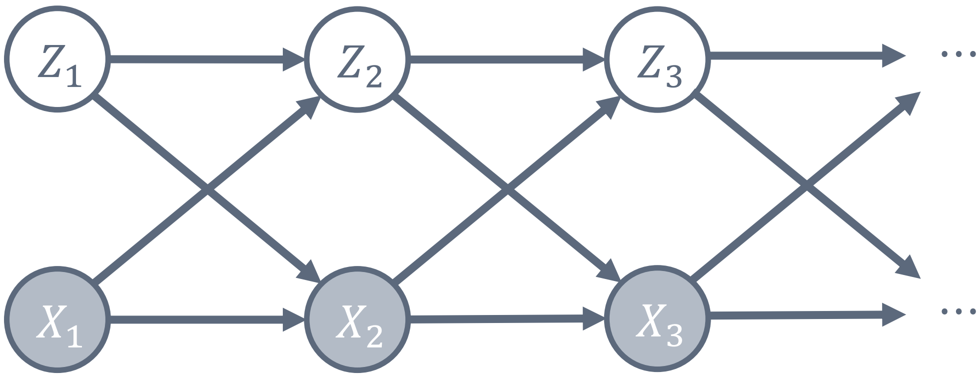
$X_{1\dots T}$ denotes the stochastic process $X_1, \dots, X_T$

Truth $q(X_{1\dots T})$

Model $p(X_{1\dots T}|w)$

Minimize $I_{q\|p}(X_{1\dots T})$

Mutable Processes



Discriminative $q(Z_{1\ldots T}, X_{1\ldots T}) = q(Z_{1\ldots T} | X_{1\ldots T}) q(X_{1\ldots T})$

Generative $p(Z_{1\ldots T}, X_{1\ldots T})$

Constraints on mutable process $q(Z_{1\ldots T} | X_{1\ldots T})$ affect optimal value of upper bound $I_{q\|p}(Z_{1\ldots T}, X_{1\ldots T})$

Computational Costs

Free Process

- No constraints on mutable process $q(Z_{1\dots T}|X_{1\dots T})$

$$q(Z_{1\dots T}|X_{1\dots T}) = q(Z_1|X_{1\dots T})\, q(Z_2|Z_1, X_{1\dots T}) \cdots q\left(Z_T|Z_{1\dots(T-1)}, X_{1\dots T}\right)$$

- By chain rule, optimal value of $I_{q\|p}(Z_{1\dots T}, X_{1\dots T})$ is

$$I_{\text{free}} = I_{q\|p}(X_{1\dots T})$$

Computational Costs

Online Learning

- Given past observations $X_{1\ldots k}$, mutable variable Z_{k+1} is independent of present and future observations $X_{(k+1)\ldots T}$

$$q(Z_{k+1}|Z_{1\ldots k}, X_{1\ldots T}) = q(Z_{k+1}|Z_{1\ldots k}, X_{1\ldots k})$$

- Optimal value I_{online} of $I_{q||p}(Z_{1\ldots T}, X_{1\ldots T})$ under constraints; *cost of online learning* is $I_{\text{online}} - I_{\text{free}}$

Computational Costs

Limited Memory

- Mutable variables Z_{k+1} are Markov, with access only to latest memory Z_k and observation X_k

$$q(Z_{k+1}|Z_{1\ldots k}, X_{1\ldots k}) = q(Z_{k+1}|Z_k, X_k)$$

- Optimal value I_{mem} of $I_{q\|p}(Z_{1\ldots T}, X_{1\ldots T})$ under constraints; *cost of limited memory* is $I_{\text{mem}} - I_{\text{online}}$

Computational Costs

Limited Sensing

- Each $X_k = (V_k, U_k)$ where mutable process observes only V_k and generative process fixes distribution of U_k

$$q(Z_{k+1}|Z_k, V_k, U_k) = q(Z_{k+1}|Z_k, V_k)$$

- Assume true process with hidden variables is Markov

- Optimal value I_{sense} of $I_{q||p}(Z_{1\ldots T}, X_{1\ldots T})$ under constraints; *cost of limited sensing* is $I_{sense} - I_{mem}$

Stationarity

Assume q has unique stationary distribution $\bar{\pi}$ (which holds under mild ergodicity conditions)

Let $\bar{q}$ be Markov process with initial distribution $\bar{\pi}$ but same transition probabilities as q.

Under above constraints on mutable process,

$$\lim_{T\to\infty} \frac{1}{T} I_{q\|p}(Z_{1\ldots T}, X_{1\ldots T})$$
$$= \lim_{n\to\infty} I_{q\|p}(Z_{n+1}, X_{n+1} | Z_n, X_n)$$
$$= I_{\bar{q}\|p}(Z_2, X_2 | Z_1, X_1).$$

Online Learning Algorithm

Assume parametric $q_\lambda(Z_{1\ldots T}|X_{1\ldots T})$ and $p_\theta(Z_{1\ldots T}, X_{1\ldots T})$

Goal. Minimize $I_{\bar{q}\|p}(Z_2, X_2|Z_1, X_1)$ over λ, θ

Strategy. Variational inference, stochastic approximation

1. Sample X_{n+1} from true process $q(X_{n+1}|X_n)$
2. Sample Z_{n+1} from mutable process $q_\lambda(Z_{n+1}|Z_n, X_n)$
3. Sample $\nabla_\theta I_{\bar{q}\|p}(Z_2, X_2|Z_1, X_1)$ using Z_{n+1}, X_{n+1}
4. Sample $\nabla_\lambda I_{\bar{q}\|p}(Z_2, X_2|Z_1, X_1)$ using Z_{n+1}, X_{n+1}
5. Update λ, θ and repeat until convergence

Gradients

(easy part, similar to training fully-observed model)

$$\nabla_\theta I_{\bar{q}||p}(Z_2, X_2 | Z_1, X_1)$$
$$= \mathbb{E}_{\bar{q}}[\nabla_\theta \log p_\theta(Z_2, X_2 | Z_1, X_1)]$$
$$= \lim_{T\to\infty} \mathbb{E}_{q(Z_{1\ldots T}, X_{1\ldots T})}[\nabla_\theta \log p_\theta(Z_T, X_T | Z_{T-1}, X_{T-1})]$$

(hard part, involves derivative under stationary distribution)

$$\nabla_\lambda I_{\bar{q}||p}(Z_2, X_2 | Z_1, X_1)$$
$$= \lim_{T\to\infty} \mathbb{E}_{q(Z_{1\ldots T}, X_{1\ldots T})}\left[\left(\log \frac{q_\lambda(Z_T, X_T | Z_{T-1}, X_{T-1})}{p_\theta(Z_T, X_T | Z_{T-1}, X_{T-1})}\right) \times \sum_{t=1}^{T-1} \nabla_\lambda \log q_\lambda(Z_{t+1} | Z_t, X_t)\right]$$

Stochastic Approximation

$$X_{n+1} \sim q(X_{n+1}|X_n)$$
$$Z_{n+1} \sim q_{\lambda_n}(Z_{n+1}|Z_n, X_n)$$
$$\theta_{n+1} = \theta_n + \eta_{n+1} \nabla_\theta \log p_\theta(Z_{n+1}, X_{n+1}|Z_n, X_n)|_{\theta=\theta_n}$$
$$\alpha_{n+1} = \alpha_n + \nabla_\lambda \log q_\lambda(Z_{n+1}|Z_n, X_n)|_{\lambda=\lambda_n}$$
$$\gamma_{n+1} = \xi(X_{n+1}|X_n) + \log \frac{q_{\lambda_n}(Z_{n+1}|Z_n, X_n)}{p_{\theta_n}(Z_{n+1}, X_{n+1}|Z_n, X_n)}$$
$$\lambda_{n+1} = \lambda_n - \eta_{n+1}\alpha_{n+1}\gamma_{n+1}$$

Proof of Convergence

$$X_{n+1} \sim q(X_{n+1}|X_n)$$

$$Z_{n+1} \sim q_{\lambda_n}(Z_{n+1}|Z_n, X_n)$$

$$\theta_{n+1} = \theta_n + \eta_{n+1}\nabla_\theta \log p_\theta(Z_{n+1}, X_{n+1}|Z_n, X_n)|_{\theta=\theta_n}$$

$$\alpha_{n+1} = \boldsymbol{\rho}\, \alpha_n + \nabla_\lambda \log q_\lambda(Z_{n+1}|Z_n, X_n)|_{\lambda=\lambda_n}$$

$$\gamma_{n+1} = \xi(X_{n+1}|X_n) + \log \frac{q_{\lambda_n}(Z_{n+1}|Z_n, X_n)}{p_{\theta_n}(Z_{n+1}, X_{n+1}|Z_n, X_n)}$$

$$\lambda_{n+1} = \lambda_n - \eta_{n+1}\alpha_{n+1}\gamma_{n+1}$$

- Convergence requires discount factor $0 < \rho < 1$
- Proof involves theory of *biased stochastic approximation*

Karimi, Belhal, Blazej Miasojedow, Eric Moulines, and Hoi-To Wai. "Non-asymptotic analysis of biased stochastic approximation scheme." PMLR, 2019.

Exploration and Exploitation

By assumption, Z_k independent of X_k given their past, so

$$I_{\bar{q}\|p}(Z_2, X_2 | Z_1, X_1) = I_{\bar{q}\|p}(Z_2 | Z_1, X_1) + I_{\bar{q}\|p}(X_2 | Z_1, X_1)$$

exploitation exploration

Exploitation. $I_{\bar{q}\|p}(Z_2 | Z_1, X_1)$ minimized when $q(Z_2 | Z_1, X_1)$ equals/exploits $p(Z_2 | Z_1, X_1)$ from the generative process.

Exploration. $I_{\bar{q}\|p}(X_2 | Z_1, X_1)$ minimized when $p(X_2 | Z_1, X_1)$ close to true $q(X_2 | X_1)$, where Z_1 controlled by stationary distribution of $q(Z_2, X_2 | Z_1, X_1)$. During optimization, Z_1 that help predict the next observation is explored and preferred.

Exploration and Exploitation

$$I_{\bar{q}\|p}(Z_2, X_2 | Z_1, X_1) = I_{\bar{q}\|p}(Z_2 | Z_1, X_1) + I_{\bar{q}\|p}(X_2 | Z_1, X_1)$$

exploitation exploration

Exploitative Modulation

$$\alpha_{n+1}\left(\log q_{\lambda_n}(Z_{n+1} | Z_n, X_n) - \log p_{\theta_n}(Z_{n+1} | Z_n, X_n)\right)$$

Explorative Modulation

$$\alpha_{n+1}\left(\xi(X_{n+1} | X_n) - \log p_{\theta_n}(Z_{n+1} | Z_n, X_n)\right)$$

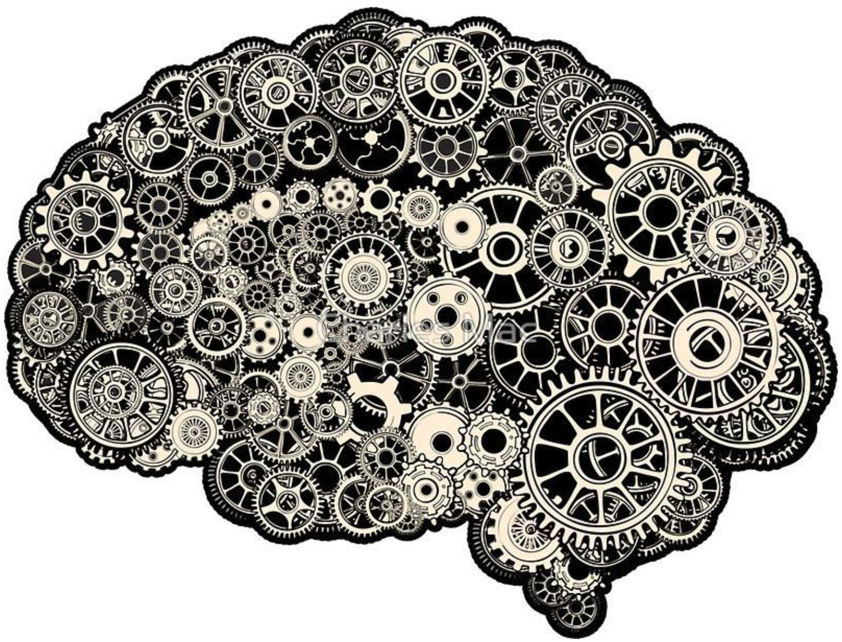
novelty

For convergence, function $\xi(X_{n+1} | X_n)$ can be any estimate of the true $\log q(X_{n+1} | X_n)$.

Conclusions

- Singularities of relative information determine asymptotic behavior of learning algorithms
- Variational inference is a powerful framework for designing learning algorithms and analyzing tradeoffs
- To design and train recurrent networks, we need both discriminative and generative processes
- Stationarity of discriminative process affects exploitation, exploration and convergence

Questions?



https://shaoweilin.github.io/