Outline

1. Likelihood Models and Variational Inference
2. Generalized Evidence Bounds
3. Experimental Results
4. Summary and future work
Introduction

Probabilistic formulation of generative modelling

- Capturing the underlying uncertainty of data
  - accounting natural variations, observational noise, etc.
  - leveraging Baye rules for inference
Likelihood and Divergence

Statistical evidence of observation

- Likelihood (density) $p(x)$: how likely $x$ occurs
  - $p(x) \geq 0$ and $\int p(x) \, dx = 1$

Kullback-Leibler (KL) divergence

- Assessing the difference between distributions $p(x)$ and $q(x)$
- $\mathbb{D}_{KL}(p \parallel q) := \mathbb{E}_{X \sim p}[\log p(X) - \log q(X)]$
  - $\mathbb{D}_{KL}(p \parallel q) \geq 0$
  - $\mathbb{D}_{KL}(p \parallel q) \neq \mathbb{D}_{KL}(q \parallel p)$
  - $\mathbb{D}_{KL}(p \parallel q) = 0$ iff $p(x) \overset{a.s.}{=} q(x)$
Likelihood and Divergence

Maximum likelihood criteria

- \( \{x_i\}_{i=1}^n \) are independent observations of true distribution \( p_d(x) \)
- \( \{p_\alpha(x)\}_{\alpha \in A} \) is a family of distributions parameterized by \( \alpha \)
- Log-likelihood loss: \( \hat{\ell}(\alpha) = \sum_i \log p_\alpha(x_i) \)
- Maximizing \( \ell(\alpha) \) \iff\ Minimizing \( D_{KL}(p_d \parallel p_\alpha) \)
  - Maximum likelihood estimation (MLE): \( \alpha^* = \arg \max_{\alpha \in A} \{\ell(\alpha)\} \)
  - e.g. regression, classification, etc.

Maximum likelihood criteria
Latent Variable Models

Joint distribution and conditional distribution

- \( p(x, z) = p(x|z)p(z) = p(x)p(z|x) \)
  - \( x \in \mathbb{R}^p: \) data (observable)
  - \( z \in \mathbb{R}^d: \) latent variable (unobservable, to be inferred)

- Bayes posterior

\[
\log p(z|x) = \log \frac{p(x, z)}{p(x)} = \log p(x|z) + \log p(z) - \log p(x)
\]

- Likelihood
- Prior
- Evidence

- Sample from the posterior: MH, HMC, Langevin, etc.
- More recently w/ quasi Monte-Carlo: Stein (SVGD)
Categorization of Likelihood-based Approaches

Likelihood-based models

Maximizing likelihood $\leftrightarrow$ minimizing KL-divergence ($KL(p_d \parallel p_G)$)

- Explicit density model (Direct approach)
  - Direct specification of a likelihood model $p(x; \theta)$
  - e.g. Autoregressive density estimators (NADE, MADE, MAF, IAF, etc.)

- Latent variable model (Indirect approach)
  - More flexible formulation through latent variables $p(x; \theta) = \int p(x|z; \theta)p(z) \, dz$
  - Indirect optimization wrt some variational bounds / surrogates
  - e.g. Variational inference (VI), expectation maximization (EM), restricted Boltzmann machine (RBM), etc.

Likelihood model trained with continuous normalizing flow (T Chen, 2018).
Variational bound of the likelihood

\[ \log p(x) = \mathbb{E}_{Z \sim q(z)} \left[ \log p(x, Z) - \log q(Z) \right] + \text{KL}(q(z) \parallel p(z|x)) \]

- **Evidence Lower BOund (ELBO)**
- **Posterior gap**

- \( q(z) \) is called *approximate posterior*
- ELBO \( \rightarrow \log p(x) \) as \( \text{KL}(q(z) \parallel p(z|x)) \rightarrow 0 \)
- ELBO \( \leq \log p(x) \) as KL-divergence is non-negative

Variational inference (VI)

- \( \{q_\beta(z|x)\}_{\beta \in \mathcal{B}} \) family of approximate posterior distributions
- \( \alpha^\dagger, \beta^\dagger = \arg \max_{(\alpha, \beta) \in \mathcal{A} \times \mathcal{B}} \text{ELBO}(p_\alpha(x, z), q_\beta(z|x)) \)
Evidence Lower Bound (ELBO)

ELBO as prior regularization

\[
\text{ELBO}(p(x, z), q(z)) = \mathbb{E}_{Z \sim q(z)}[\log p(x|Z)] - \text{KL}(q(z) \parallel p(z))
\]

- Conditional likelihood
- Prior regularization

Choice of posteriors

- Mean field approximation: \( q(z) = \prod_{k=1}^{d} q_k(z_k) \)
- The variational principle
  \[
  q_k^*(z_k) \propto \exp(\mathbb{E}_{Z_{-k} \sim q_{-k}^*} [\log p(x|Z_{-k}, z_k) + \log p(z_k)])
  \]
- More tractable yet often over-simplified
Recent Advances in Variational Inference

Tightening the bound

- Flexible likelihoods
  - Flexible posteriors, contrastive likelihoods, adaptive prior, etc.
- Alternative variational objectives
  - $\beta$-VAE, IW-VAE, Renyi-VI, PBBVI, etc.

Better optimization

- Variance reduction
  - VIMCO, RELAX, Stick-landing, etc.
- Update schemes
  - Natural gradient descent, proximal descent, gradient flows, etc.
A Motivating Example for Generalizing ELBO

Model equivalence under MLE: How to select?

- Finite sample can lead to model evidence tie
  - Simultaneous occurrence of over-fitting and under-fitting on the empirical samples
  - Reweight the samples to encourage better fitting of low evidence samples (under fitted)

Illustration. \( \log \)-evidence tie

Both models have the same expected \( \log \)-likelihood on the empirical samples. However, *Model-2* is arguably more favorable as the model-likelihoods are more balanced.
A Maximal Entropy Argument for Model Selection

Intuition

- Consider a discrete approximation $q(x)$ to $p_d(x)$
  - samples are only allowed to take values from the empirical distribution $\hat{p}_n := \frac{1}{n} \sum \delta(x_i)$
- A natural choice $\hat{q}_m$ yielding a good (discrete) approximation would be $\hat{q}_m(x_i) \propto q_{\beta^*}(x_i)$, where $q_{\beta^*}(x)$ is a best fit for $p_d(x)$
- $\min \text{KL}(\hat{q}_m \parallel \hat{p}_m) \iff \max \{- \sum_i q_{\beta}(x_i) \log q_{\beta}(x_i)\}$ (entropy)

Max-Ent principle for model selection

- Under the constraints of testable information, the most likely model encodes largest uncertainty
Surrogate objective

- **(Intuitively)** We assign **different weights** wrt the model evidence under current model to each sample during training, *encouraging better fitting for low evidence samples.*

- **(Formally)** We **rig** the evidence function. 😁

---

*Poster #215 C Tao*  
Variational Inference and Model Selection with GLBO
Alternative Motivations for Rigging the Evidence Function

Log-evidence function and Jensen inequality

- Likelihood $p(x)$ can vary on very different scales
  - Extremely large, difficult to optimize
  - The log (evidence) transform took them to a comparable scale

- Jensen inequality
  - Let $f(x)$ be a concave function, then $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$.
  - The gap is smaller if $f(x)$ is less curved.

Intuition

If we can reduced the concavity of evidence function, then better variational bounds can be expected.
Generalized Evidence Lower Bound (GLBO)

Assumptions

- $\phi(u)$ is concave
  - New evidence function that encodes our belief to encourage more concentrated log-evidence distribution.
  - The assumption on concavity is technical
- $\psi(u)$ is convex and monotonically increasing (invertible)
- $h(u) := \psi(\phi(u))$ concave
  - Annealing the convexity (curvature) of evidence function
Generalized Evidence Lower Bound (GLBO)

**Generalized Evidence Lower Bound**

\[
\text{GLBO}(x; K) = \psi^{-1} \left( \mathbb{E}_{Z_k \sim q} \left[ h \left( \frac{1}{K} \sum_{k=1}^{K} \frac{p(x, Z_k)}{q(Z_k|x)} \right) \right] \right)
\]

\{Z_k\}_{k=1}^{K} \text{ are } K \text{ i.i.d. samples from } q(z|x).

**Properties of GLBO**

- **Theorem 1.** \(\text{GLBO}(x; 1) \leq \text{GLBO}(x; 2) \leq \ldots \xrightarrow{K \to \infty} \phi(p(x))\)

- **Theorem 2.** If \(\phi(u) = \log u\), \(\text{GLBO}(x; K) \geq \text{ELBO}(x; K)\).

\[
\text{ELBO}(x; K) = \mathbb{E}_{Z_k \sim q} \left[ \log \left( \frac{1}{K} \sum_{k} \frac{p(x, Z_k)}{q(Z_k|x)} \right) \right]
\]

(Importance weighted evidence lower bound)
Generalized Evidence Lower Bound (GLBO)

Special case: $\chi$ evidence lower bound (CLBO)

- $\text{CLBO}(x; K, T) := \text{GLBO}(x; K, \phi = \log(u), \psi = \exp(\frac{1}{T}u))$

- **Theorem 3.** CLBO is monotonic wrt $T$, and

  $\text{ELBO}(x; K) \overset{T \to \infty}{\leftarrow} \text{CLBO}(x; K, T) \overset{T \to 1}{\rightarrow} \log p(x)$

  *Related to Rényi bounds.*

Variance trade-off

- When $T$ is sufficiently large,

  $\text{CLBO}(x; 1, T) \approx \text{ELBO} + \text{var}[f(x, Z)]/(2T)$,

  where $f(x, z) := \log p_\alpha(x, z) - \log q_\beta(z|x)$.

  *Tighter bounds are associated with more volatile estimates.*
Generalized Evidence Lower Bound (GLBO) Demo

Figure: Comparison of theoretical bounds on the toy distribution.
Upper bounds

- Upper bounds can be similarly established, constructing a sandwich formula for the generalized evidence. However, they are *less useful* due to excess variance associated.

Comparison to other model selection criteria

- Many information-theoretic model selection criteria takes the form *model complexity + expected model evidence*.
- Instead, we consider the distribution of model evidence.
**Results**

**Table:** Average test log-likelihood on MNIST. † results collected from (burda2015importance, li2016renyi)

<table>
<thead>
<tr>
<th>L</th>
<th>K</th>
<th>VAE†</th>
<th>IW-VAE†</th>
<th>Renyi†</th>
<th>GLBO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>-86.47</td>
<td>-85.41</td>
<td>-85.42</td>
<td>-84.71</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-86.35</td>
<td>-84.80</td>
<td>-84.81</td>
<td>-84.30</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>-85.01</td>
<td>-83.92</td>
<td>-84.04</td>
<td>-83.45</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>-84.75</td>
<td>-83.12</td>
<td>-83.44</td>
<td>-82.94</td>
</tr>
</tbody>
</table>
Table: ELBO, AIS and reconstruction error on MNIST for models with flexible posterior distributions. ‡ results collected from (mescheder2017adversarial).

<table>
<thead>
<tr>
<th>Model</th>
<th>ELBO</th>
<th>AIS</th>
<th>Recon. Err.</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVB+GLBO</td>
<td>-79.97±0.15</td>
<td>-81.2</td>
<td>57.2±0.12</td>
</tr>
<tr>
<td>AVB‡</td>
<td>-82.7±0.2</td>
<td>-81.7</td>
<td>57.±0.2</td>
</tr>
<tr>
<td>VAE‡</td>
<td>-85.7±0.2</td>
<td>-81.9</td>
<td>59.4±0.2</td>
</tr>
<tr>
<td>AuxiliaryVAE‡</td>
<td>-85.6±0.2</td>
<td>-81.6</td>
<td>59.6±0.2</td>
</tr>
<tr>
<td>VAE/IAF‡</td>
<td>-85.5±0.2</td>
<td>-82.1</td>
<td>59.6±0.2</td>
</tr>
</tbody>
</table>
## Results

Table: Bayesian regression & classification on UCI datasets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Test log likelihood (higher is better)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VI</td>
</tr>
<tr>
<td>Boston</td>
<td>-2.90 ± .07</td>
</tr>
<tr>
<td>Concrete</td>
<td>-3.39 ± .02</td>
</tr>
<tr>
<td>Energy</td>
<td>-2.39 ± .03</td>
</tr>
<tr>
<td>Kin8nm</td>
<td>0.90 ± .01</td>
</tr>
<tr>
<td>Naval</td>
<td>3.73 ± .12</td>
</tr>
<tr>
<td>CCPP</td>
<td>-2.89 ± .02</td>
</tr>
<tr>
<td>Winequality</td>
<td>-0.98 ± .01</td>
</tr>
<tr>
<td>Yacht</td>
<td>-3.43 ± .16</td>
</tr>
<tr>
<td>Protein</td>
<td>-2.99 ± .01</td>
</tr>
</tbody>
</table>
Summary and Outlook

Contributions of GLBO

- Tighter evidence bounds generalizing prior arts
- New model selection criteria and optimizations schemes

On-going and future research

- Integrating likelihood-based and likelihood-free methods constructs an interesting research direction
- Harnessing detrimental variances to improve training
Thank you.

Welcome to our poster #215 @ Hall B tonight.

Please also visit poster #113, #192 in case we are on rotating duty. ^_^
Bibliography

- T Raiforth, et al. Tighter variational bounds are not necessarily better. *ICML 2018.*
Additional Results

Poster #215  C Tao  Variational Inference and Model Selection with GLBO