Variational Inference with Tail-adaptive $f$-Divergence

Dilin Wang  Hao Liu  Qiang Liu

Department of Computer Science
The University of Texas at Austin

TEXAS
The University of Texas at Austin
Overview

- **Variational inference (VI):** a key algorithmic engine for Bayesian (deep) learning, generative models, reinforcement learning, etc.

- **Key challenge:** Under-estimate uncertainty for multi-modal distributions.

- **This Work:** VI based on a new tail-adaptive $f$-divergence.
  - Efficiently promoting **mass-covering** without scarifying **robustness**.
  - Yields a very **simple** algorithm.
Variational Inference: KL Divergence

- Approximate intractable distributions by minimizing KL divergence:

\[
\min_{\theta} \text{KL}(q_{\theta} \mid\mid p) = \mathbb{E}_{q_{\theta}} \left[ \log \left( \frac{q_{\theta}(x)}{p(x)} \right) \right].
\]

- \( p \): given complex distribution (e.g., the poster distribution in Bayes inference).
- \( q_{\theta} \): approximation family (e.g., Gaussian, mixtures, normalizing flow).
Variational Inference: $\alpha$-Divergence

More generally, $\alpha$-divergence:

$$\min_{\theta} D_{\alpha}(p \parallel q_\theta) = \frac{1}{\alpha(\alpha - 1)} \mathbb{E}_{q_\theta} \left[ \left( \frac{p(x)}{q_\theta(x)} \right)^\alpha - 1 \right].$$

Large values ($\alpha \geq 1$) promote mass-covering.

See [Minka+ 01,05; Hernandez-Lobato+ 16; Li & Turner 16; Opper & Winther 05; Burda+ 16; Cotter+15; Dieng+ 16; Capper+ 18; Ryu & Boyd 14].
Stochastic Optimization

- Gradient with reparameterization trick [e.g., Li & Turner 16]:

\[
\nabla_\theta D_\alpha(p \parallel q_\theta) = -\frac{1}{\alpha(\alpha - 1)} \mathbb{E}_{q_\theta} \left[ \left( \frac{p(x)}{q_\theta(x)} \right)^\alpha \nabla_\theta g_\theta(\epsilon) \nabla_x \log \left( \frac{p(x)}{q_\theta(x)} \right) \right],
\]

where we assume \( x \sim q_\theta \) by \( x = g_\theta(\xi) \), \( \xi \sim q_0 \).

- \( \alpha = 0 \implies \) standard VI (with reparameterization trick).
- Large \( \alpha \implies \) promoting mode-covering.

Unfortunately, large \( \alpha \) causes

- Large or infinite variance.
- Or even infinite mean:

\[
D_\alpha(p \parallel q_\theta) = +\infty, \quad \text{and} \quad \nabla_\theta D_\alpha(p \parallel q_\theta) = +\infty
\]

objective and gradient are ill-defined!
Stochastic Optimization

- Gradient with reparameterization trick [e.g., Li & Turner 16]:

\[
\nabla_{\theta} D_{\alpha}(p \parallel q_{\theta}) = -\frac{1}{\alpha(\alpha - 1)} \mathbb{E}_{q_{\theta}} \left[ \left( \frac{p(x)}{q_{\theta}(x)} \right)^\alpha \nabla_{\theta} g_{\theta}(\epsilon) \nabla_{x} \log \left( \frac{p(x)}{q_{\theta}(x)} \right) \right],
\]

where we assume \(x \sim q_{\theta}\) by \(x = g_{\theta}(\xi), \xi \sim q_{0}\).
- \(\alpha = 0 \implies\) standard VI (with reparameterization trick).
- Large \(\alpha \implies\) promoting mode-covering.

Unfortunately, large \(\alpha\) causes
- Large or infinite variance.
- Or even infinite mean:

\[
D_{\alpha}(p \parallel q_{\theta}) = +\infty, \quad \text{and} \quad \nabla_{\theta} D_{\alpha}(p \parallel q_{\theta}) = +\infty
\]

objective and gradient are ill-defined!
Infinite Mean/Variance is not Uncommon!

- **Gaussian Example:** \( p = \mathcal{N}(0, \sigma_p^2) \) and \( q = \mathcal{N}(0, \sigma_q^2) \).
- When \( q \) has smaller variance than \( p \) (\( \sigma_q^2 < \sigma_p^2 \)), define

\[
\alpha_* = \frac{\sigma_p^2}{\sigma_p^2 - \sigma_q^2},
\]

then \( \alpha \geq \alpha_* \implies \) infinite mean, \( \mathbb{E}_q[(p/q)^\alpha] = +\infty \);
\( \alpha \geq \alpha_*/2 \implies \) infinite variance, \( \text{var}_q[(p/q)^\alpha] = +\infty \).
Fat-Tailed Distribution and Tail Index

- Density ratio \( w := \frac{p(x)}{q_\theta(x)} \), \( x \sim q_\theta \) is often fat-tailed, whose tail probability decays like a power law:
  \[
  T(t) := \Pr(w \geq t) \approx t^{-\alpha_*}.
  \]

- \( \alpha_* \) is called tail index; it defines the upper limit of finite power moments:
  \[
  \mathbb{E}[w^\alpha] < \infty \quad \text{iff} \quad \alpha < \alpha_*.
  \]

- \( \alpha_* \) defines “strongest mode-covering”, but:
  - \( \alpha_* \) is unknown.
  - \( \alpha_* \) changes across iterations when \( q_\theta \) varies.
  - \( \alpha_* \) is difficult to estimate; see [Hill+ 75; Vehtari+ 15]].
Easier to estimate tail probability $T(t)$, instead of $\alpha_\ast$.

Note the tail probability $T(t) = \Pr(w \geq t) \approx t^{-\alpha_\ast}$. We have

$$T(t)^{-1} \approx t^{\alpha_\ast}.$$ 

Estimating $T(t)$ is simple and easy: with sample $W := \{w_1, \ldots, w_n\}$,

$$\hat{T}(t) = \sum_{i=1}^{n} \frac{\mathbb{I}(w_i \geq t)}{n} = \frac{\text{rank}_W(t)}{n}$$

We can use $\hat{T}(t)^{-1}$ to replace $t^{\alpha_\ast}$ to achieve tail-adaptive $\alpha_\ast$-divergence.
**f-Divergence**

- **f-divergence**: for strictly \textit{convex} function \( f \),

\[
D_f(p \parallel q_\theta) = \mathbb{E}_{q_\theta} \left[ f \left( \frac{p(x)}{q_\theta(x)} \right) - f(1) \right]
\]

- \textbf{Gradient}:

\[
\nabla_\theta D_f(p \parallel q_\theta) = -\mathbb{E}_{q_\theta} \left[ \gamma \left( \frac{p(x)}{q_\theta(x)} \right) \nabla_\theta g_\theta(\epsilon) \nabla_x \log \left( \frac{p(x)}{q_\theta(x)} \right) \right],
\]

where \( \gamma(t) := f''(t)t^2 \) (positive).

- \( \alpha \)-divergence: \( f(t) = t^\alpha / (\alpha(\alpha - 1)) \Rightarrow \gamma_f(t) = t^\alpha \).

\textbf{Key:} Positive \( \gamma(t) \iff \text{strictly convex } f(t) \).

\textbf{We can define } \( f \)-\textit{divergence using any positive function } \( \gamma(t) \)!
Tail-adaptive $f$-Divergence

- Specifying $f$-divergence by any non-negative function $\gamma$:

$$\nabla_\theta D_f(p \parallel q_\theta) = -E_{q_\theta} \left[ \gamma \left( \frac{p(x)}{q_\theta(x)} \right) \nabla_\theta g_\theta(\epsilon) \nabla_x \log \left( \frac{p(x)}{q_\theta(x)} \right) \right],$$

- **Standard VI**: $\gamma(t) = 1$
- **$\alpha$-divergence**: $\gamma(t) = t^\alpha$
- **Tail-adaptive $f$-divergence**: We take

$$\gamma(t) = \hat{T}(t)^{-\beta} = \left( \frac{n}{\text{rank}_W(t)} \right)^\beta \approx t^{\alpha*\beta}, \quad \beta \in [0, 1]$$

- Implicitly define a safe $f$-divergence; **guaranteed finite mean**.
- $f$ adaptively changes with the tail of $p/q_\theta$.
- Always achieves **strongest possible mode-covering**.
Toy Gaussian Mixture Models (GMM)

- $p$: Randomly generated 10D GMMs.
- $q_\theta$: GMM with more components than $p$. 

![Graph](image_url)

- $\alpha = 0$
- $\alpha = 0.5$
- $\alpha = 1$
- Ours

Gaussian $\rightarrow$ highly multimodal

$\alpha$: Tail-adaptive $f$-divergence
## Bayes Neural Nets on Benchmarks

<table>
<thead>
<tr>
<th>Dataset</th>
<th>ours</th>
<th>$\alpha = 0$</th>
<th>$\alpha = 0.5$</th>
<th>$\alpha = 1.0$</th>
<th>$\alpha = 2.0$</th>
<th>$\alpha = +\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boston</td>
<td>2.828</td>
<td>2.956</td>
<td>2.990</td>
<td>2.937</td>
<td>2.981</td>
<td>2.985</td>
</tr>
<tr>
<td>Concrete</td>
<td>5.371</td>
<td>5.592</td>
<td>5.381</td>
<td>5.462</td>
<td>5.499</td>
<td>5.481</td>
</tr>
<tr>
<td>Energy</td>
<td>1.377</td>
<td>1.431</td>
<td>1.531</td>
<td>1.413</td>
<td>1.458</td>
<td>1.458</td>
</tr>
<tr>
<td>Kin8nm</td>
<td>0.085</td>
<td>0.088</td>
<td><strong>0.083</strong></td>
<td>0.084</td>
<td>0.084</td>
<td><strong>0.083</strong></td>
</tr>
<tr>
<td>Naval</td>
<td><strong>0.001</strong></td>
<td><strong>0.001</strong></td>
<td><strong>0.004</strong></td>
<td>0.005</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Wine</td>
<td>0.636</td>
<td>0.634</td>
<td>0.634</td>
<td><strong>0.633</strong></td>
<td>0.635</td>
<td>0.634</td>
</tr>
<tr>
<td>Yacht</td>
<td><strong>0.849</strong></td>
<td>0.861</td>
<td>1.146</td>
<td>1.221</td>
<td>1.222</td>
<td>1.234</td>
</tr>
</tbody>
</table>

**Table:** Average test RMSE
Improving Soft Actor Critic (SAC) [Haarnoja el al., 18]

Our method used to **improve policy optimization** in SAC.

- **Ant**
- **HalfCheetah**
- **Humanoid(rllab)**
- **Walker**
- **Hopper**

![Graphs showing improvements in different environments](image-url)
Conclusion

Proposed a new tail-adaptive $f$-divergence

1. Strong mass-covering properties
2. Easy to implement; no hyper-parameters
3. Outperforms standard VI based on KL/$\alpha$-divergence
4. Easily extended to score-function gradient

Thank You