Adversarial Variational Bayes: Unifying Variational Autoencoders and Generative Adversarial Networks (ICML 2017)

Lars Mescheder\textsuperscript{1}, Sebastian Nowozin \textsuperscript{2}, Andreas Geiger\textsuperscript{3}

\textsuperscript{1}Autonomous Vision Group, MPI Tubingen & \textsuperscript{2}Microsoft Research Cambridge & \textsuperscript{3}Computer Vision and Geometry Group, ETH Zurich

Presented by Paidamoyo Chapfuwa
July 7, 2017
Main Idea

Maximize Marginal log-likelihood: \( \log p_\theta(x^1, \ldots, x^N) = \sum_{i=1}^{N} \log p_\theta(x^i) \)

\[
\log p_\theta(x^i) = \log \int p_\theta(x^i, z) dz = \log \int p_\theta(x^i, z) \frac{q_\phi(z|x^i)}{q_\phi(z|x^i)} dz \geq -D_{KL}(q_\phi(z|x^i)||p_\theta(z)) + \mathbb{E}_{q_\phi(z|x^i)}[\log p_\theta(x^i|z)]
\]

Improves VAE inference with GAN via:

1. VAE generative model.
2. Adversarial inference model: only need to sample from \( q_\phi(z|x^i) \).

Figure: Vanilla VAE
Reformulate ELBO

- KL regularization in VAE replaced with Adversarial score.
- Noise injected to input data.

\[
\max_{\theta} \max_{\phi} \mathbb{E}_{p_{D}(x)} \mathbb{E}_{q_{\phi}(z|x)} \left[ \log p(z) - \log q_{\phi}(z|x) + \log p_{\theta}(x|z) \right]
\]

log-likelihood ratio

(2)

Figure: VAE
Adversarial Strategy

- $T(x, z)$ discriminates pairs $(x, z)$ of $p_D(x)p(z)$ from $p_D(x)q_\phi(z|x)$.

\[
\max_T \mathbb{E}_{p_D(x)}\mathbb{E}_{q_\phi(z|x)}[\log \sigma(T(x, z))] + \mathbb{E}_{p_D(x)}\mathbb{E}_{p(z)}[\log(1 - \sigma(T(x, z)))]
\]

(3)

- Replace log-likelihood ratio from (2) with $T^*(x, z)$.

**Theorem 1: optimal discriminator $T^*$**

- Assuming fixed $p_\theta(x|z)$ and $q_\phi(z|x)$.

\[
T^*(x, z) = \log q_\phi(z|x) - \log p(z)
\]

(4)

- The function $t \rightarrow a \log(t) + b \log(1 - t)$ attains maximum at $t = \frac{a}{a+b}$.
Reparameterize s.t: \( \nabla_\phi \mathbb{E}_{q_\phi(z)}[\log p_\theta(x|z)] = \mathbb{E}_\epsilon \nabla_\phi [\log p_\theta(x|z_\phi(x, \epsilon))] \)

Replace log-likelihood ratio with \( T^*(x, z_\phi(x, \epsilon)) \).

\[
\max_\theta \max_\phi \mathbb{E}_{p_D(x)} \mathbb{E}_\epsilon [-T^*(x, z_\phi(x, \epsilon)) + \log p_\theta(x|z_\phi(x, \epsilon))] \quad (5)
\]

Discriminator network \( T \) optimized independent of \( \phi \)

**Theorem 2:** gradient with respect to \( \phi \) in \( T^*(x, z) \) not necessary

\[
\mathbb{E}_{q_\phi(z|x)}[\nabla_\phi T^*(x, z)] = 0 \quad (6)
\]

From Theorem 1: \( \mathbb{E}_{q_\phi(z|x)}[\nabla_\phi \log q_\phi(z)] = \int q_\phi(z) \frac{\nabla_\phi(z)q_\phi(z)}{q_\phi(z)} = 0 \)
Theoretical Nash-equilibrium

**Theorem 3: Game theory point of view**

- If \((\theta^*, \phi^*, T^*)\) defines a Nash-equilibrium for the game then:

\[
T^*(x, z) = \log q_{\phi^*}(z|x) - \log p(z) \tag{7}
\]

1. \(\theta^*\) is a maximum-likelihood assignment
2. \(q_{\phi^*}(z|x)\) is equal to true posterior \(p_{\theta^*}(z|x)\)
3. \(T^*\) is the pointwise mutual information between \(x\) and \(z\)

\[
T^*(x, z) = \log \frac{p_{\theta^*}(x, z)}{p_{\theta^*}(x)p(z)} \tag{8}
\]
Adaptive Contrast (AC)

Challenge: In practice $T(x, z)$ fails to be close to the optimal $T^*(x, z)$

- [Friedman’01]: Logistic regression works best when comparing 2 similar distributions.
- Introduce $r_\alpha(z|x)$ to approximate $q_\phi(z|x)$, s.t mean and variances match.
  - i.e $r_\alpha(z|x) \sim N(0, I)$; normalize $q_\phi(z|x)$ s.t $\tilde{z} = \frac{z-\mu(x)}{\sigma(x)}$

$$\max_{\theta} \max_{\phi} \mathbb{E}_{p_D(x)} \mathbb{E}_{q_\phi(z|x)} \left[ -T^*(x, z) - \log r_\alpha(z|x) + \log p_\theta(x, z) \right]$$ (9)

* $T^*(x, z)$ is optimal discriminator distinguishing samples from $r_\alpha(z|x)$ and $q_\phi(z|x)$
Algorithm 1 Adversarial Variational Bayes (AVB)

1: \( i \leftarrow 0 \)
2: **while** not converged **do**
3: \( \text{Sample} \ \{x^{(1)}, \ldots, x^{(m)}\} \text{ from data distrib. } p_D(x) \)
4: \( \text{Sample} \ \{z^{(1)}, \ldots, z^{(m)}\} \text{ from prior } p(z) \)
5: \( \text{Sample} \ \{\epsilon^{(1)}, \ldots, \epsilon^{(m)}\} \text{ from } \mathcal{N}(0, 1) \)
6: Compute \( \theta \)-gradient (eq. 3.7):
\[
g_\theta \leftarrow \frac{1}{m} \sum_{k=1}^{m} \nabla_{\theta} \log p_{\theta}(x^{(k)} | z_{\phi}(x^{(k)}, \epsilon^{(k)}))
\]
7: Compute \( \phi \)-gradient (eq. 3.7):
\[
g_\phi \leftarrow \frac{1}{m} \sum_{k=1}^{m} \nabla_{\phi} \left[ -T_\psi(x^{(k)}, z_{\phi}(x^{(k)}, \epsilon^{(k)})) \right. \\
+ \log p_{\theta}(x^{(k)} | z_{\phi}(x^{(k)}, \epsilon^{(k)})) \left. \right]
\]
8: Compute \( \psi \)-gradient (eq. 3.3):
\[
g_\psi \leftarrow \frac{1}{m} \sum_{k=1}^{m} \nabla_{\psi} \left[ \log (\sigma(T_\psi(x^{(k)}, z_{\phi}(x^{(k)}, \epsilon^{(k)})))) \right. \\
+ \log (1 - \sigma(T_\psi(x^{(k)}, z^{(k)}))) \left. \right]
\]
9: Perform SGD-updates for \( \theta, \phi \) and \( \psi \):
\[
\theta \leftarrow \theta + h_i \ g_\theta, \quad \phi \leftarrow \phi + h_i \ g_\phi, \quad \psi \leftarrow \psi + h_i \ g_\psi
\]
10: \( i \leftarrow i + 1 \)
11: **end while**

* Crucial to keep \( T_\psi \) network close to optimal while optimizing.
Experiments: Eight Schools

- Measure coaching effects $y_i, i = 1, \ldots, 8; y_i \sim N(\mu + \theta \cdot \eta_i, \sigma_i)$
- Placed a $N(0, 1)$ prior on all model parameters $(\mu, \tau, \eta_i)$.

**Figure:** AVB vs. VB where HMC is the ground truth

**Figure:** KL divergence over number of iterations
Experiments: Toy Data

<table>
<thead>
<tr>
<th>Metric</th>
<th>VAE</th>
<th>AVB</th>
</tr>
</thead>
<tbody>
<tr>
<td>log-likelihood</td>
<td>-1.568</td>
<td>-1.403</td>
</tr>
<tr>
<td>reconstruction error</td>
<td>$88.5 \cdot 10^{-3}$</td>
<td>$5.77 \cdot 10^{-3}$</td>
</tr>
<tr>
<td>ELBO</td>
<td>-1.697</td>
<td>$\approx -1.421$</td>
</tr>
<tr>
<td>$\text{KL}(q_\phi(z), p(z))$</td>
<td>$\approx 0.165$</td>
<td>$\approx 0.026$</td>
</tr>
</tbody>
</table>

**Figure:** Optimal log-likelihood score $= -\log 4 \approx -1.386$

**Figure:** $4(2 \times 2)$ binary images for Training

**Figure:** Distribution of latent code

(a) VAE
(b) AVB
Experiments: MNIST

<table>
<thead>
<tr>
<th>Method</th>
<th>$\log p(x) \geq$</th>
<th>$\log p(x) \approx$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AVB (8-dim)</td>
<td>$-83.6 \pm 0.4$</td>
<td>$-91.2 \pm 0.6$</td>
</tr>
<tr>
<td>AVB + AC (8-dim)</td>
<td>$-96.3 \pm 0.4$</td>
<td>$-89.6 \pm 0.6$</td>
</tr>
<tr>
<td>AVB + AC (32-dim)</td>
<td>$-79.5 \pm 0.3$</td>
<td>$-80.2 \pm 0.4$</td>
</tr>
<tr>
<td>VAE (8-dim)</td>
<td>$-98.1 \pm 0.5$</td>
<td>$-90.9 \pm 0.6$</td>
</tr>
<tr>
<td>VAE (32-dim)</td>
<td>$-87.2 \pm 0.3$</td>
<td>$-81.9 \pm 0.4$</td>
</tr>
<tr>
<td>VAE + NIF (T=80)</td>
<td>$-85.1$</td>
<td>$-$</td>
</tr>
<tr>
<td>VAE + HVI (T=16)</td>
<td>$-88.3$</td>
<td>$-85.51$</td>
</tr>
<tr>
<td>convVAE + HVI (T=16)</td>
<td>$-84.1$</td>
<td>$-81.94$</td>
</tr>
<tr>
<td>VAE + VGP (2hI)</td>
<td>$-81.3$</td>
<td>$-$</td>
</tr>
<tr>
<td>DRAW + VGP</td>
<td>$-79.9$</td>
<td>$-$</td>
</tr>
<tr>
<td>VAE + IAF</td>
<td>$-80.8$</td>
<td>$-79.10$</td>
</tr>
<tr>
<td>Auxiliary VAE (L=2)</td>
<td>$-83.0$</td>
<td>$-$</td>
</tr>
</tbody>
</table>

(Rezende & Mohamed, 2015)
(Salimans et al., 2015)
(Salimans et al., 2015)
(Tran et al., 2015)
(Tran et al., 2015)
(Kingma et al., 2016)
(Maaløe et al., 2016)

Figure: Log-Likelihoods on binarized MNIST AVB vs. methods improving on VAEs

(a) Training data          (b) Random samples

Figure: Generated random samples
Experiments: CelebA

Figure: AVB without AC: Generated random samples
Experiments: CelebA

Figure: AVB without AC: Reconstruction + Interpolation