Doubly Semi-Implicit Variational Inference

Dmitry Molchanov[1,2,*], *Valery Kharitonov*[2,*], Artem Sobolev[1], Dmitry Vetrov[1,2]

[1] Samsung AI Center in Moscow
[2] Samsung-HSE Lab

October 19, 2018

Overview

Variational Inference and Variational Learning

- **Variational Inference**
  Given a joint $p(x, z) = p(x \mid z)p(z)$, find the posterior $p(z \mid x)$:

$$\mathcal{L} = \mathbb{E}_{p(x)}\mathbb{E}_{q_\phi(z \mid x)} \log \frac{p(x \mid z)p(z)}{q_\phi(z \mid x)} \to \max_{\phi}.$$

- **Variational Learning**
  Approximately maximize the marginal log-likelihood $\log p(x \mid \theta, \chi)$

$$\mathbb{E}_{p(x)} \log p(x \mid \theta, \chi) \geq \mathbb{E}_{p(x)}\mathbb{E}_{q_\phi(z \mid x)} \log \frac{p_\theta(x \mid z)p_\chi(z)}{q_\phi(z \mid x)} \to \max_{\phi,\theta,\chi}.$$

(Semi-)implicit distributions

- We call a distribution *implicit*, if we can sample from it, but there is no closed-form density available.
- In particular, we call a distribution $q_\phi(z)$ *semi-implicit*, if it can be represented as following:

$$q_\phi(z) = \int q_\phi(z \,|\, \psi) q_\phi(\psi) \, \mathrm{d}\psi,$$

  where $q_\phi(z \,|\, \psi)$ has analytically tractable density and both $q_\phi(z \,|\, \psi)$ and $q_\phi(\psi)$ are reparameterizable.
- Any implicit distribution can be approximated with a semi-implicit distribution arbitrarily well:

$$q_\phi(z) \approx \int \mathcal{N}(z \,|\, z', \sigma^2) q_\phi(z') \, \mathrm{d}z', \quad \sigma^2 \to 0.$$

Semi-implicit distributions

- If we can efficiently sample from $q_\phi(\psi)$, it is easy to sample from $q_\phi(z)$:
$$\psi \sim q_\phi(\psi), \quad z \sim q_\phi(z \mid \psi).$$
- For example, samples from $q_\phi(\psi)$ could be the output of some neural network $\psi = f(\varepsilon, \phi)$, where ε is a sample from non-parametric noise distribution and ϕ are parameters of the NN.

Recap: methods for implicit variational inference

- Discriminator-based density ratio estimation
- Approaches based on reverse models
  - Hierarchical variational inference
  - Unbiased implicit variational inference
- Denoising-based inference
- Other approaches:
  - (D)SIVI ((doubly) semi-implicit VI)
  - KIVI (kernel implicit VI)
  - OPVI (operator VI)
  - ...

Semi-Implicit Variational Inference

- How do we perform variational inference with a semi-implicit approximate posterior $q_\phi(z)$?
- Basic idea: estimate the marginal density using Monte Carlo:

$$\begin{aligned} q_\phi(z) &= \int q_\phi(z \mid \psi) q_\phi(\psi) \, \mathrm{d}\psi \\ &\approx \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid \psi^k), \quad \psi^k \sim q_\phi(\psi). \end{aligned}$$

SIVI upper bound

- We can obtain an upper bound on ELBO by plugging in this estimate (apply Jensen's inequality):

$$\overline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi(z \mid \psi^0)} \left[\log p(x, z) - \log \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid \psi^k) \right]$$
$$\geqslant \mathbb{E}_{q_\phi(z)} \left[\log p(x, z) - \log q_\phi(z) \right] = \mathcal{L}.$$

- Additionally, this upper bound is asymptotically exact:

$$\lim_{K \to \infty} \overline{\mathcal{L}}_K^q = \mathcal{L}.$$

- It is also monotonic:

$$\overline{\mathcal{L}}_K^q \geqslant \overline{\mathcal{L}}_{K+1}^q \geqslant \mathcal{L}.$$

- Unfortunately, we cannot use an upper bound to maximize ELBO!

SIVI lower bound

- Lower bound is quite similar (the proof is not as simple though):

$$\underline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi(z \mid \psi^0)} \left[\log p(x, z) - \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \psi^k) \right]$$
$$\leqslant \mathbb{E}_{q_\phi(z)} [\log p(x, z) - \log q_\phi(z)] = \mathcal{L}.$$

- The lower bound is also asymptotically exact:

$$\lim_{K \to \infty} \underline{\mathcal{L}}_K^q = \mathcal{L}.$$

- It monotonic as well:

$$\underline{\mathcal{L}}_K^q \leqslant \underline{\mathcal{L}}_{K+1}^q \leqslant \mathcal{L}.$$

- Use as a surrogate to maximize ELBO.

SIVI lower bound intuition

- It turns out, we can rewrite $\underline{\mathcal{L}}_K^q$ as follows:

$$\underline{\mathcal{L}}_K^q = \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi(z|\psi^0)} \left[\log p(x,z) - \log \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \psi^k) \right]$$
$$= \mathbb{E}_{\psi^{0..K} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi^K(z|\psi^{0..K})} \left[\log p(x,z) - \log q_\phi^K(z|\psi^{0..K}) \right],$$

  where

$$q_\phi^K(z|\psi^{0..K}) := \frac{1}{K+1} \sum_{k=0}^{K} q_\phi(z \mid \psi^k).$$

- Note that the second expectation in the second expression is different.
- $\underline{\mathcal{L}}_K^q$ is actually the average of the true ELBOs in the finite mixture approximation model, averaged over all such mixtures.

SIVI sandwich

- Since we have both a lower bound and an upper bound on the ELBO, we can use them together to evaluate how well we approximate the ELBO on convergence.
- Furthermore, we can use this sandwich to evaluate not only models that are trained using SIVI objective, but *any* semi-implicit model.

Variational Learning with Semi-Implicit Priors

- We can use a similar idea to come up with a lower bound on ELBO in case of semi-implicit *priors*.
- For now, let's assume that $q_\phi(z)$ is explicit and

$$p_\theta(z) = \int p_\theta(z|\zeta) p_\theta(\zeta)\, \mathrm{d}\zeta.$$

- Again, using Jensen's inequality,

$$\underline{\mathcal{L}}_K^p = \mathbb{E}_{\zeta^{1..K} \sim p_\theta(\zeta)} \mathbb{E}_{z \sim q_\phi(z)} \left[\log \frac{p(x|z)}{q_\phi(z)} + \log \frac{1}{K} \sum_{k=1}^{K} p_\theta(z \mid \zeta^k) \right]$$
$$\leqslant \mathbb{E}_{z \sim q_\phi(z)} \left[\log \frac{p(x|z)}{q_\phi(z)} + \log p_\theta(z) \right] = \mathcal{L}.$$

- Use the lower bound as a surrogate for ELBO optimization.

Variational Learning with Semi-Implicit Priors

- Again, the presented lower bound is monotonic w.r.t. K and asymptotically exact.
- Unfortunately, we cannot derive an upper bound for the case of implicit priors using the same technique.
- We have to resort to the variational representation to bound the KL-divergence between $q_\phi(z)$ and $p_\theta(z)$ from below:

$$
\begin{aligned}
\mathrm{KL}(q_\phi(z) \| p_\theta(z)) = 1 + \sup_{g:\mathrm{dom}\, z \to \mathbb{R}} \left\{ \mathbb{E}_{p_\theta(z)} g(z) - \mathbb{E}_{q_\phi(z)} e^{g(z)} \right\} \geqslant \\
\geqslant 1 + \sup_{\eta} \left\{ \mathbb{E}_{p_\theta(z)} g_\eta(z) - \mathbb{E}_{q_\phi(z)} e^{g_\eta(z)} \right\}
\end{aligned}
$$

Doubly Semi-Implicit Variational Inference

- We can combine the two lower bounds to obtain an objective for variational inference and learning when both the prior and the approximate posterior are semi-implicit:

$$
\begin{aligned}
\underline{\underline{\mathcal{L}}}^{q,p}_{K_1,K_2} = \mathbb{E}_{z \sim q_\phi(z)} \log p(x \mid z) - \\
- \mathbb{E}_{\psi^{0..K_1} \sim q_\phi(\psi)} \mathbb{E}_{z \sim q_\phi(z \mid \psi^0)} \log \frac{1}{K_1} \sum_{k=0}^{K_1} q_\phi(z \mid \psi^k) + \\
+ \mathbb{E}_{\zeta^{1..K_2} \sim p_\theta(\zeta)} \mathbb{E}_{z \sim q_\phi(z)} \log \frac{1}{K_2} \sum_{k=1}^{K_2} p_\theta(z \mid \zeta^k).
\end{aligned}
$$

- Use variational representation of KL divergence to obtain an upper bound:

$$
\overline{\mathcal{L}}^{q,p}_{\eta} = \mathbb{E}_{q_\phi(z)} \log p(x \mid z) - \mathbb{E}_{p_\theta(z)} g(z, \eta) + \mathbb{E}_{q_\phi(z)} e^{g(z,\eta)} - 1.
$$

Applications

- Variational inference with hierarchical priors and posteriors;
- VAE with semi-implicit priors and posteriors;
- Deep Weight Prior;
- Incremental learning.

Variational inference with hierarchical priors

- It is common to specify a hyperprior over the hyperparameters of the prior in discriminative models:

$$p(t, w, \alpha \mid x) = p(t \mid w, x) p(w \mid \alpha) p(\alpha).$$

- Usually, we approximate the joint posterior:

$$q_\phi(w, \alpha) \approx p(w, \alpha \mid X_{tr}, T_{tr}).$$

- Then, for prediction, we use the marginal approximate posterior:

$$\begin{aligned} p(t \mid x, X_{tr}, T_{tr}) &= \int p(t \mid x, w) p(w \mid X_{tr}, T_{tr}) \, dw = \\ &= \int p(t \mid x, w) \int p(w, \alpha \mid X_{tr}, T_{tr}) \, d\alpha \, dw \approx \\ &\approx \int p(t \mid x, w) \int q_\phi(w, \alpha) d\alpha \, dw = \\ &= \int p(t \mid x, w) q_\phi(w) \, dw. \end{aligned}$$

Variational inference with hierarchical priors

- The objective for the joint inference is the following:

$$\mathcal{L}^{joint}(\phi) = \mathbb{E}_{q_\phi(w,\alpha)} \log \frac{p(t \mid x, w) p(w \mid \alpha) p(\alpha)}{q_\phi(w, \alpha)}.$$

- We are actually not interested in the joint posterior, since we marginalize out the hyperparameters anyway.
- Idea: optimize for the marginal posterior directly using semi-implicit prior $p(w) = \int p(w \mid \alpha) p(\alpha) \, \mathrm{d}\alpha$ and posterior $q_\phi(w) = \int q_\phi(w \mid \alpha) q_\phi(\alpha) \, \mathrm{d}\alpha$:

$$\mathcal{L}^{marginal}(\phi) = \mathbb{E}_{q_\phi(w)} \log \frac{p(t \mid x, w) p(w)}{q_\phi(w)}.$$

Variational inference with hierarchical priors

$$\mathcal{L}^{joint}(\phi) = \mathbb{E}_{q_\phi(w,\alpha)} \log \frac{p(t \mid x, w) p(w \mid \alpha) p(\alpha)}{q_\phi(w, \alpha)}.$$

$$\mathcal{L}^{marginal}(\phi) = \mathbb{E}_{q_\phi(w)} \log \frac{p(t \mid x, w) p(w)}{q_\phi(w)}.$$

Theorem

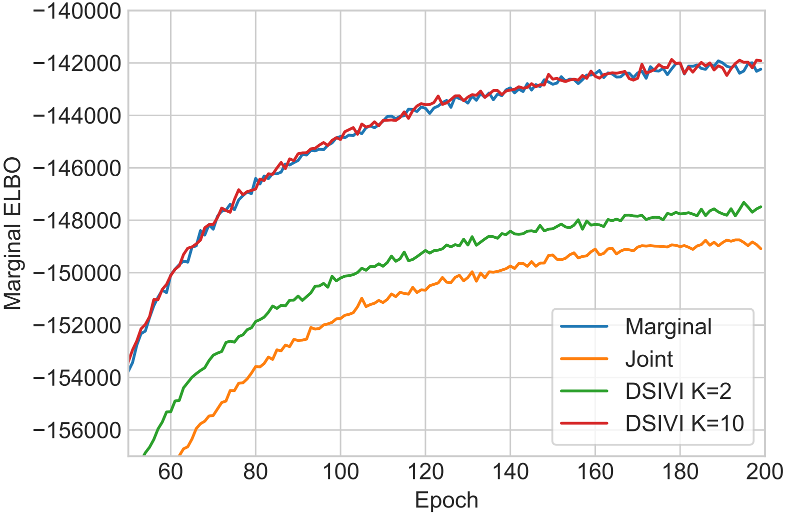
Let ϕ_j and ϕ_m maximize $\mathcal{L}^{joint}$ and $\mathcal{L}^{marginal}$ correspondingly. Then

$$\mathrm{KL}(q_{\phi_m}(w) \,\|\, p(w \mid X_{tr}, T_{tr})) \le \mathrm{KL}(q_{\phi_j}(w) \,\|\, p(w \mid X_{tr}, T_{tr})).$$

Variational inference with hierarchical priors

- Let $p(w \mid \alpha) = \mathcal{N}(w \mid 0, \alpha^{-1})$, $p(\alpha) = \mathrm{Gamma}(\alpha \mid a = 0.5, b = 2)$. Then $p(w) = \mathrm{t}(w \mid \nu = 1)$.
- Take $q_\phi(w, \alpha) = q_\phi(w) q_\phi(\alpha)$, where $q_\phi(w)$ is factorized normal and $q_\phi(\alpha)$ is factorized log-normal.
- $p(t \mid x, w)$ is a fully connected NN on MNIST.
- Consider 3 ways to perform approximate inference: joint inference over (w, α), marginal inference over w with Student's t-prior, and inference over w with a semi-implicit prior.

Variational inference with hierarchical priors



VAE with semi-implicit posteriors

- SIVI allows one to perform variational inference in VAEs with multiple stochastic layers in the encoder.

$$\ell_t = T_t(\ell_{t-1}, \varepsilon_t, x; \phi), \quad \varepsilon_t \sim q_t(\varepsilon), t = 1 \dots M,$$
$$\mu(x, \phi) = f(\ell_M, x; \phi), \quad \Sigma(x, \phi) = g(\ell_M, x; \phi);$$
$$q_\phi(z \,|\, x, \mu, \Sigma) = \mathcal{N}(z \,|\, \mu(x, \phi), \Sigma(x, \phi))$$

- Unlike the regular VAE, μ and σ are now random variables (cf. ψ in $q_\phi(z \,|\, \psi)$).
- The marginal $q_\phi(z \,|\, x)$ is thus more expressive than a fully-factorized Gaussian.

VAE with semi-implicit posteriors

Table 2. Comparison of the negative log evidence between various algorithms.

Methods	$-\log p(\boldsymbol{x})$
Results below form Burda et al. (2015)	
VAE + IWAE	$= 86.76$
IWAE + IWAE	$= 84.78$
Results below form Salimans et al. (2015)	
DLGM + HVI (1 leapfrog step)	$= 88.08$
DLGM + HVI (4 leapfrog step)	$= 86.40$
DLGM + HVI (8 leapfrog steps)	$= 85.51$
Results below form Rezende & Mohamed (2015)	
DLGM+NICE (Dinh et al., 2014) (k = 80)	≤ 87.2
DLGM+NF (k = 40)	≤ 85.7
DLGM+NF (k = 80)	≤ 85.1
Results below form Gregor et al. (2015)	
DLGM	≈ 86.60
NADE	$= 88.33$
DBM 2hl	≈ 84.62
DBN 2hl	≈ 84.55
EoNADE-5 2hl (128 orderings)	$= 84.68$
DARN 1hl	≈ 84.13
Results below form Maaløe et al. (2016)	
Auxiliary VAE (L=1, IW=1)	≤ 84.59
Results below form Mescheder et al. (2017)	
VAE + IAF (Kingma et al., 2016)	$\approx 84.9 \pm 0.3$
Auxiliary VAE (Maaløe et al., 2016)	$\approx 83.8 \pm 0.3$
AVB + AC	$\approx 83.7 \pm 0.3$
SIVI (3 stochastic layers)	$= 84.07$
SIVI (3 stochastic layers)+ IW($\tilde{K} = 10$)	$= 83.25$

VAE with semi-implicit priors

- It can be shown that the so-called aggregated posterior distribution is the optimal prior distribution for a VAE in terms of the value of ELBO:

$$p^*(z) = \frac{1}{N} \sum_{n=1}^{N} q_\phi(z \mid x_n),$$

- When N is large, such prior can lead to overfitting and it is highly computationally inefficient.
- Middle ground (VampPrior and VampPrior-data):

$$p^{Vamp}(z) = \frac{1}{K} \sum_{k=1}^{K} q_\phi(z \mid u_k),$$

where u_k are either trainable inducing imputs (VampPrior) or pre-sampled images from the dataset (VampPrior-data).

VAE with semi-implicit priors

- There are two ways to improve upon VampPrior.
- We can regard the aggregated posterior as a semi-implicit distribution:

$$p^*(z) = \frac{1}{N}\sum_{n=1}^{N} q_\phi(z|x_n) = \int q_\phi(z|x) p_{data}(x)\, \mathrm{d}x.$$

- Alternatively, we may consider an arbitrary learnable semi-implicit distribution as a prior:

$$p_\theta^{SI}(z) = \int p_\theta(z \mid \zeta) p_\theta(\zeta)\, \mathrm{d}\zeta.$$

VAE with semi-implicit priors

Table: We compare VampPrior with its semi-implicit modifications, DSIVI-agg and DSIVI-prior. We report the the IWAE objective $\mathcal{L}^S$ for VampPrior-data, and the corresponding lower bound $\underline{\mathcal{L}}_K^{p,S}$ for DSIVI-based methods. Only the prior distribution is semi-implicit.

Method	LL
VAE+VampPrior-data	-85.05
VAE+VampPrior	-82.38
VAE+DSIVI-prior (K=2000)	≥ -82.27
VAE+DSIVI-agg (K=500)	≥ -83.02
VAE+DSIVI-agg (K=5000)	$\geq -\mathbf{82.16}$
HVAE+VampPrior-data	-81.71
HVAE+VampPrior	-81.24
HVAE+DSIVI-agg (K=5000)	$\geq -\mathbf{81.09}$

References

M. Yin and M. Zhou (2018)
Semi-implicit variational inference
ICML volume 80, pages 5660–5669.

Molchanov, D., Kharitonov, V., Sobolev, A. and Vetrov, D. (2018)
Doubly Semi-Implicit Variational Inference
arXiv preprint arXiv:1810.02789

Tomczak, J.M. and Welling, M. (2018).
VAE with a VampPrior
AISTATS, pages 1214–1223

IWAE and IW-DSIVAE bound

- IWAE bound:

$$\log p(x) \geq \mathcal{L}^S = \mathbb{E}_{z^{1..S} \sim q_\phi(z)} \log \frac{1}{S} \sum_{i=1}^{S} \frac{p(x \mid z^i) p(z^i)}{q_\phi(z_i \mid x)}$$

- IW-DSIVAE bound:

$$\underline{\mathcal{L}}^{q,p,S}_{K_1,K_2} = \mathbb{E}_{\psi^{1..K_1} \sim q_\phi(\psi)} \mathbb{E}_{\zeta^{1..K_2} \sim p_\theta(\zeta)} \Bigg[\mathbb{E}_{(z^1,\hat{\psi}^1),\dots,(z^S,\hat{\psi}^S) \sim q_\phi(z,\psi)} \Bigg[\log \frac{1}{S} \sum_{i=1}^{S} \frac{p(x \mid z^i) \frac{1}{K_2} \sum_{k=1}^{K_2} p_\theta(z^i \mid \zeta^k)}{\frac{1}{K_1+1} (q_\phi(z^i \mid \hat{\psi}^i) + \sum_{k=1}^{K_1} q_\phi(z^i \mid \psi^k))} \Bigg] \Bigg].$$