Perturbation Theory for Variational Inference

Manfred Opper[1], Marco Fraccaro[2], Ulrich Paquet[3], Alex Susemihl[1] and Ole Winther[2]

[1]Technical University of Berlin [2]Technical University of Denmark [3]Apple

Overview

- We view the variational free energy and its accompanying evidence lower bound as a first-order term from a perturbation of the true log partition function and derive a power series of corrections.
- This allows for example to get better estimates of normalizing constants or to correct predictions in latent variable models whose permutation-invariant parameters are self-pruned and ignored by the variational approximation (e.g. matrix factorization).

Variational Inference

Given a random variable x with probability distribution $p(x) = \frac{\mu(x)e^{-H(x)}}{Z}$, we often need to compute

$$-\log Z = -\log \int \mu(x)\, e^{-H(x)} dx \qquad \text{or} \qquad E[f(x)] = \int f(x)\, \mu(x) e^{-H(x)} dx \,.$$

If these are not analytically tractable, we introduce a *variational approximation* $q(x) = \frac{1}{Z_q}\mu(x)e^{-H_q(x)}$ whose parameters are found minimizing the free energy

$$F[q] = KL[q||p] - \log Z = E_q\left[\log \frac{q(x)}{p(x)}\right] - \log Z = E_q[V(x)] - \log Z_q \,,$$

where we have defined $V(x) = H(x) - H_q(x)$.

Perturbative corrections

Perturbation theory aims at finding approximate solutions to a problem given exact solutions of a simpler related sub-problem (the VB solution in our case).

Normalizing constant. By defining $\widehat{H}_\lambda = H_q + \lambda V = (1-\lambda)H_q + \lambda H$ (notice that this gives $\widehat{H}_1 = H$ and $\widehat{H}_0 = H_q$), we can write

$$-\log \int \mu(x)e^{-\widehat{H}_\lambda(x)}dx = -\log Z_q - \log E_q[e^{-\lambda V(x)}]$$

$$= \underbrace{-\log Z_q + \lambda E_q[V]}_{F[q] \text{ for } \lambda=1} - \frac{\lambda^2}{2}E_q\left[(V - E_q[V])^2\right] + \frac{\lambda^3}{3!}E_q\left[(V - E_q[V])^3\right] + \ldots$$

Knowing the VB solution, we can therefore correct our estimate of $\log Z$ using higher order terms.
Note that this may not be a convergent series, but lead to an asymptotic expansion only.

Expectations. Defining E_λ as the expectation with respect to $p_\lambda(x) = \mu(x)e^{-H_\lambda(x)}$, so that $E = E_1$ and $E_q = E_0$, we get:

$$E_\lambda[f(x)] = \frac{\int f(x)\, \mu(x)e^{-\widehat{H}_0(x) - \lambda V(x)}\, dx}{\int \mu(x)e^{-\widehat{H}_0(x)-\lambda V(x)}\, dx} = \frac{E_0\left[f(x)e^{-\lambda V(x)}\right]}{E_0\left[e^{-\lambda V(x)}\right]}$$

$$= \frac{E_0\left[f(x)\left(1 - \lambda V + \frac{\lambda^2}{2}V^2 - \frac{\lambda^3}{3!}V^3 \pm \ldots\right)\right]}{E_0\left[\left(1 - \lambda V + \frac{\lambda^2}{2}V^2 - \frac{\lambda^3}{3!}V^3 \pm \ldots\right)\right]} \overset{\frac{1}{1-z}=1+z+z^2+\ldots}{=}$$

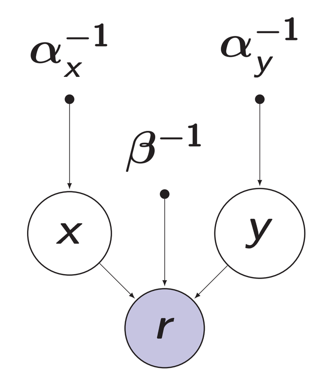
$$= \underbrace{E_0[f(x)]} - \lambda \mathrm{Cov}_0[f(x), V] - \lambda^2 E_0[V]\mathrm{Cov}_0[f(x), V] + \frac{\lambda^2}{2}\mathrm{Cov}_0[f(x), V^2] \pm \ldots$$

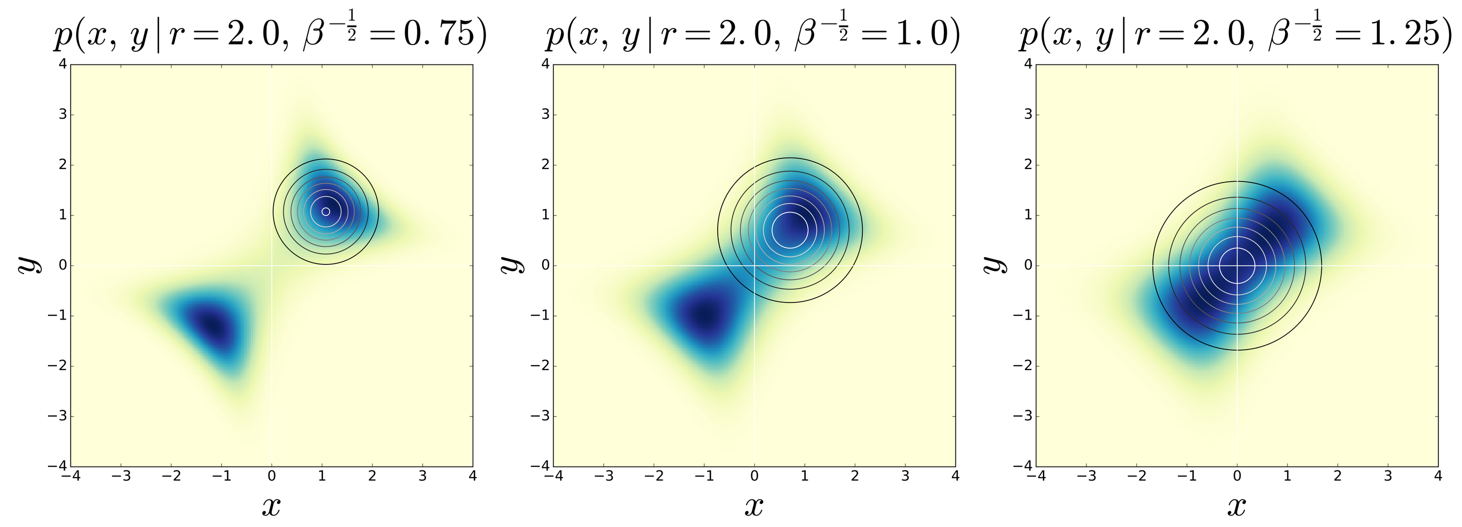
We can correct the expectation w.r.t the VB solution using higher order terms.

Example: Variational Matrix Factorization

Model:

$$r = xy + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \beta^{-1})$$
$$x \sim \mathcal{N}(0, \alpha_x^{-1})$$
$$y \sim \mathcal{N}(0, \alpha_y^{-1})$$



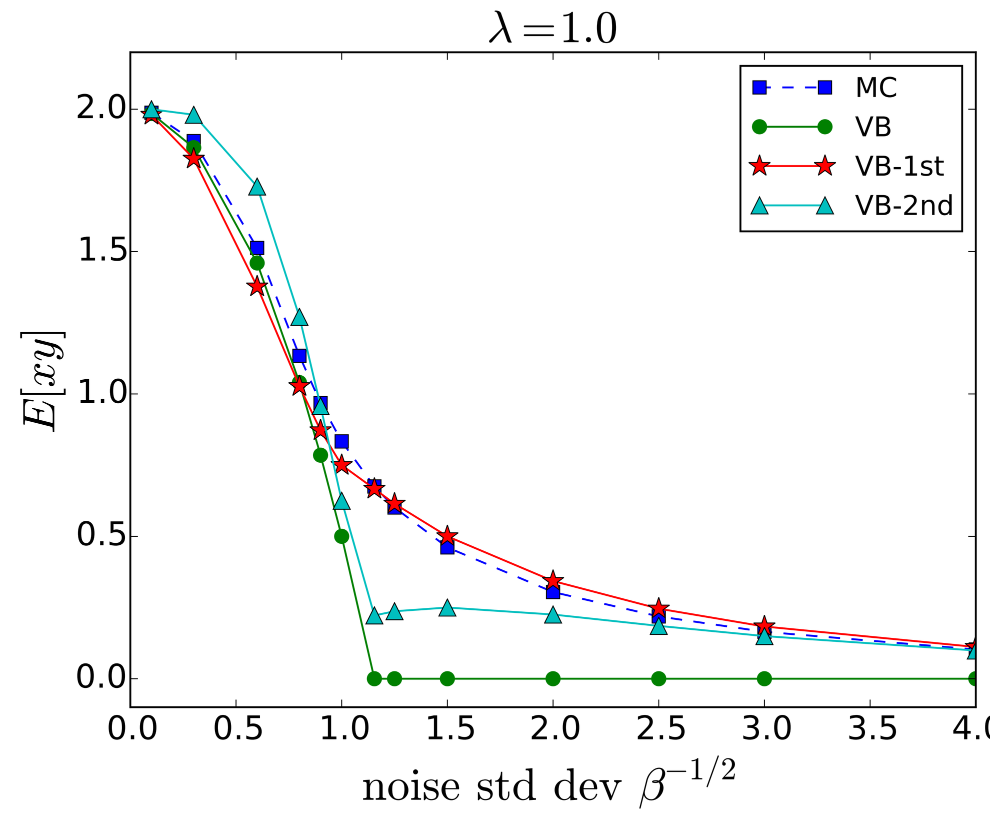


VB solution:

$$q(x) = \mathcal{N}(\mu_x, \gamma_x^{-1})\,, \quad q(y) = \mathcal{N}(\mu_y, \gamma_y^{-1})$$

Predictions + corrections:

$$H(x, y) = \frac{1}{2}\left[\beta(r - xy)^2 + \alpha_x x^2 + \alpha_y y^2\right]$$
$$H_q(x, y) = \frac{1}{2}\left[\gamma_x(x - \mu_x)^2 + \gamma_y(y - \mu_y)^2\right]$$
$$V(x, y) = H(x, y) - H_q(x, y)$$
$$E_\lambda[xy] = E_0[xy] - \lambda \mathrm{Cov}_0[xy, V(x, y)] \pm \ldots$$



Example: Variational Inference for Sparse GP Regression

- We can compute the predictive mean and covariance using a sparse approximation for the GP and then include the perturbative corrections.

$$y_i(t_i) = \underbrace{3\mathrm{sinc}(t_i)}_{x(t_i)} + \epsilon_i\,, \qquad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$x \sim GP(0, K) \text{ with RBF kernel}$$