Particle Mirror Descent

0. Reference:

Dai et al. "Scalable Bayesian Inference via Particle Mirror Descent". arXiv:1506.03101

1. Solve VFE minimization with MD

(assume we can do that exactly...)

We have some prior $p_0(\theta)$, model $p(x|\theta)$, and observation $D = \{x_1, \dots, x_N\}$.

Goal: find some tractable

$$q(\theta) \approx p(\theta|D)$$

Variational Inference: solve q by minimizing

$$\mathcal{L}(q) = KL[q(\theta)||p(\theta|D)] - \log p(D)$$

$$= KL[q(\theta)||p_0(\theta)] - \sum_{n=1}^N E_q[\log p(x_n|\theta)]$$

Variational Free Energy

(subject to $\int_\theta q(\theta) d\mu = 1$)

Stochastic approximation: $\mathcal{L}(q) = \mathbb{E}_{x\sim D}[l(q;x)]$,

$$l(q;x) = KL[q(\theta)||p_0(\theta)] - N E_q[\log p(x|\theta)]$$

Now solve it with MD!

$$q_{t+1}(\theta) = \arg\min_{\hat{q}(\theta)\in\mathcal{P}} \underbrace{\{\langle \hat{q}(\theta), \nabla l(q_t; x_t)\rangle_{L_2} + \frac{1}{\gamma_t} KL[\hat{q}||q_t]\}}_{\hat{\mathcal{L}}_t(\hat{q})}$$

where $x_t \sim D$ is the sample at time t,

γ_t the learning rate at time t,

$\mathcal{P} = \{p(\theta): \int_\theta p(\theta) d\mu = 1\}$

details:

$$g_t(\theta) \triangleq \nabla l(q_t; x_t) = \log q_t(\theta) - \log p_0(\theta) - N \log p(x_t|\theta)$$

$$0 = \nabla \hat{\mathcal{L}}_t(\hat{q}) = \nabla l(q_t; x_t) + \frac{1}{\gamma_t}[\log \hat{q}(\theta) - \log q_t(\theta)]$$

$$\Rightarrow \log q_{t+1}(\theta) \leftarrow \log q_t(\theta) - \gamma_t g_t(\theta)$$

$$\Rightarrow q_{t+1}(\theta) = \frac{q_t(\theta)\exp[-\gamma_t g_t(\theta)]}{Z}$$

$$= \frac{q_t(\theta)^{1-\gamma_t} p_0(\theta)^{\gamma_t} p(x_t|\theta)^{N\gamma_t}}{Z}$$

also called "normalised exponential gradient" if want to read more!

Particle Mirror Descent (cont.)

MD update rule:

$$q_{t+1}(\theta) \leftarrow \frac{1}{Z} q_t(\theta)^{1-\gamma_t} p_0(\theta)^{\gamma_t} p(x_t|\theta)^{N\gamma_t}, \quad x_t \sim D$$

problem 1: such q_{t+1} is generally intractable!

Solution: restrict $q \in Q$ (tractable),

compute MD update, then project it to the Q family:

$$q_{t+1}(\theta) \leftarrow \mathrm{proj}_{Q}\left[\frac{1}{Z} q_t(\theta)^{1-\gamma_t} p_0(\theta)^{\gamma_t} p(x_t|\theta)^{N\gamma_t}\right]$$

problem 2: theory of MD stochastic approximations works only for running average answer over time!

Solution: empirical verification

2. PMD with weighted particle

assume the starting proposal $\pi(\theta) \approx p(\theta|D)$:

define $q_t(\theta) = \sum_{i=1}^m \alpha_i \delta(\theta_i)$

with $\theta_i \sim \pi(\theta)$ and fixed over time

$\Rightarrow$ at time t,

$$\alpha_i \leftarrow \frac{1}{Z} \alpha_i^{1-\gamma_t} p_0(\theta_i)^{\gamma_t} p(x_t|\theta_i)^{N\gamma_t},$$

$$Z = \sum_{i=1}^m \alpha_i^{1-\gamma_t} p_0(\theta_i)^{\gamma_t} p(x_t|\theta_i)^{N\gamma_t}$$

3. PMD with weighted KDE

works when we don't have a good guess $\pi(\theta)$.

define $q_t(\theta) = \sum_{i=1}^m \alpha_i K_h(\theta - \theta_i)$, $\theta_i \sim q_{t-1}(\theta)$

$\Rightarrow$ at time t,

$$\alpha_i \leftarrow \frac{\exp[-\gamma_t g_t(\theta_i)]}{\sum_{i=1}^m \exp[-\gamma_t g_t(\theta_i)]}, \quad \theta_i \sim q_t(\theta)$$

$$= \frac{1}{Z} q_t(\theta_i)^{-\gamma_t} p_0(\theta_i)^{\gamma_t} p(x_t|\theta_i)^{N\gamma_t}$$