Reminder: Variance reduction in MC methods [See M. Hoffman's RCC slides]

$$\mathbb{E}_{p(x)}[f(x)] \approx \overbrace{\frac{1}{N}\sum_{n=1}^{N} f(x^i)}^{F} \quad , x^i \sim p(x)$$

Problems : HIGH Variance

Want : F' s.t $\quad \mathbb{E}[F'] = \mathbb{E}[F] = \mathbb{E}[f(x)]$

$$\mathrm{var}[F'] \leq \mathrm{var}[F]$$

Use $\quad g(x) \quad$ where $\quad \mu_g = \mathbb{E}[g(x)]$ is tractable

$\rightarrow \quad \mathbb{E}[f(x)] = \mathbb{E}[f(x) - g(x)] + \mu_g$

$\rightarrow$ SAME Expectation , LOWER Variance if f & g are STRONGLY CORRELATED

Now consider $\hat{f}(x) = f(x) - a\, g(x) + a\, \mu_g$

a : some constant

$$\mathrm{var}(\hat{f}) = \mathrm{var}(f) + a^2\, \mathrm{var}(g) - 2a\, \mathrm{Cov}(f, g)$$

$$\text{optimal } a = \left.\frac{\mathrm{Cov}(f, g)}{\mathrm{var}(g)}\right\} \text{ Use } \underline{\text{emperical estimate}}$$

Questions : How to choose g ?

For VI, g could be
+ another lower bound [deterministic]
+ Taylor series approximation
+ . . .

Back to VI

Consider

$$\left.\begin{array}{l} p(\theta) \\ p(y|\theta) = \prod_{n=1}^{N} p(y_n|\theta) \end{array}\right\} \Rightarrow \text{want } p(\theta|y) \propto p(\theta).p(y|\theta)$$

VI

$$F(q(\theta)) = \left\langle \log \frac{p(y|\theta).p(\theta)}{q(\theta)} \right\rangle_{q(\theta)}$$

$$= -KL(q(\theta)||p(\theta)) + \sum_{n=1}^{N} \langle \log p(y_n|\theta) \rangle_{q(\theta)}$$

However $\langle \log p(y_n|\theta) \rangle_{q(\theta)}$ ~~not~~ typically not analytically tractable

① approx. ↓ by another function, e.g. Jaakkola & Jordan 2000, Marlin + Murphy 2011

② go stochastic ...

Optimising $F(q(\theta))$ requires $\frac{dF}{d\lambda}$

$$\frac{dF}{d\lambda} = -\frac{d}{d\lambda} KL(q(\theta) \| p(\theta)) + \sum_{n=1}^{N} \underbrace{\int \frac{dq(\theta)}{d\lambda} \log p(y_n|\theta) \, d\theta}_{= \int q(\theta) . \log p(y_n|\theta) . \frac{d}{d\lambda} \log q(\theta) \, d\theta}$$

$$\approx \underbrace{-\frac{d}{d\lambda} KL(q(\theta) \| p(\theta))}_{\frac{dF_1}{d\lambda}} + \sum_{n=1}^{N} \sum_{m=1}^{M} \log p(y_n|\theta_m) \frac{d}{d\lambda} . \log q(\theta_m)$$

$$\theta_m \sim q(\theta).$$

SGD: $\lambda \leftarrow \lambda + \eta \frac{dF}{d\lambda}$

But need to lower the variance of $\log p(y_n|\theta_m) \frac{d}{d\lambda} \log q(\theta_m)$

$$\Rightarrow \frac{dF}{d\lambda} = \frac{dF_1}{d\lambda} + \sum_{n=1}^{N} \sum_{m=1}^{M} \left(\log p(y_n|\theta_m) - \hat{a} \, g_n(\theta_m) \right) \frac{d}{d\lambda} \log q(\theta_m)$$

i.e. $f \equiv \log p(y_n | \theta_m) \frac{d}{d\lambda} \log q(\theta_m)$

$g \equiv g_n(\theta_m) \cdot \frac{d}{d\lambda} \log q(\theta_m)$

$$\hat{a} = \frac{\mathrm{Cov}(f, g)}{\mathrm{var}(g)}$$

Example: Logistic regression $\log p(y_n | \theta_m) = \log \sigma(y_n x_n^T \theta_m)$

$q(\theta) = N(\theta; \mu, \Sigma)$

⊕ J & J bound $g_n(\theta) = \ln \sigma(\xi_n) + \frac{1}{2}(y_n x_n^T \theta - \xi_n) + \ldots$

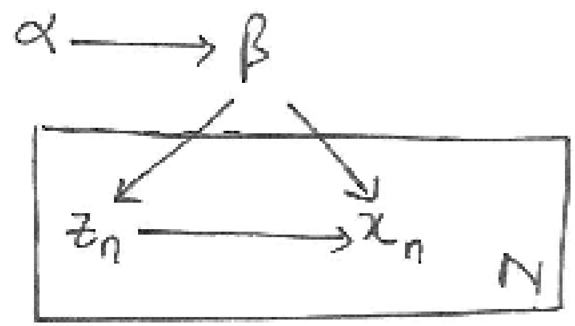
Marlin bound $g_n(\theta) =$ linear + quadratic pieces

⊕ 2nd order Taylor of $\log p(y_n | \theta_m)$ around μ:

$g_n(\theta) = \ln \sigma(y_n x_n^T \mu) + y_n (1 - \sigma_n) \cdot (\theta - \mu)^T x_n + \ldots$

Results See paper

Review of SVI (Hoffman et al)



Prob of everything.

$$p(x, z, \beta) = p(\beta) \cdot \prod_{n=1}^{N} p(x_n, z_n \mid \beta)$$

α : fixed (hyper)-params

β : global parameters

$\{z_n\}$: local parameters

$\{x_n\}$: observations

Want $\quad p(z, \beta) \;\propto\; p(x, z, \beta)$

VI $$F(q(z, \beta)) = -E_q[\log(q)] + E_q[\log p(x, z, \beta)]$$

Mean field $$q(z, \beta) = q(\beta \mid \lambda) \prod_{n=1}^{N} \prod_{j=1}^{J} q(z_{nj} \mid \phi_{nj})$$

$$\Rightarrow \left[\begin{array}{l} \lambda = E_q[\eta_g(x, z, \alpha)] \\ \phi_{nj} = E_q[\eta_\ell(x_n, z_{n,-j}, \beta)] \end{array} \right.$$

VI _ Mean field

Iterate : $t = 1:T$

$$\phi_{nj,t} \leftarrow E_{q_{t-1}}[\eta_e(x_n, z_{n,-j}, \beta)]$$

$$\lambda_t \leftarrow E_{q_{t-1}}[\eta_g(x, z, \alpha)]$$

SVI _ Mean field

Iterate $t = 1:T$

$$\phi_{nj} \leftarrow E_q[\eta_e(x_n, z_{n,-j}, \beta)]$$

$$\lambda_t \leftarrow (1-\rho_t)\lambda_{t-1} + \rho_t \hat{\lambda}_t$$

$$\hat{\lambda}_t = E_{q_{t-1}}[\eta_g(\underbrace{x_m, z_m}_{\text{mini batch}}, \alpha)]$$

$$\lambda_t = (1-\rho_t)\lambda_{t-1} + \rho_t \hat{\lambda}_t$$

Kalman Variational Filter (KVF)

$$\begin{array}{ccc} \lambda_{t-1}^{VB} & \longrightarrow & \lambda_t^{VB} \\ \downarrow & & \downarrow \\ \hat{\lambda}_{t-1} & & \hat{\lambda}_t \end{array}$$

ie observe $\hat{\lambda}_t$: noisy params

infer λ_t^{VB} : true params

More on $\lambda_t = (1-\rho_t)\cdot\lambda_{t-1} + \rho_t \hat{\lambda}_t$

$$E\left[\lambda_t - \lambda_t^{VB}\right] = (1-\rho_t)\cdot(\lambda_{t-1} - \lambda_t^{VB}) \quad \text{as } E[\hat{\lambda}_t] = \lambda_t^{VB}$$

$$\mathrm{Var}[\lambda_t] = \rho_t^2 \, \mathrm{Var}[\hat{\lambda}_t]$$

KVF automatically handles bias/variance trade off + step-size

---

$$\begin{tikzcd}
\lambda_t^{VB} \arrow[r] \arrow[d] & \lambda_{t+1}^{VB} \arrow[d] \\
\hat{\lambda}_t & \hat{\lambda}_{t+1}
\end{tikzcd}$$

$$p(\lambda_{t+1}^{VB} \mid \lambda_t^{VB}) = N(\lambda_{t+1}^{VB}; \lambda_t^{VB}, Q)$$

$$p(\hat{\lambda}_t \mid \lambda_t^{VB}) = N(\hat{\lambda}_t; \lambda_t^{VB}, R)$$

$\Rightarrow$ Every SVI iteration, perform Kalman update

$$\mu_{t|t} = (1-\rho_t)\,\mu_{t-1|t-1} + \rho_t \hat{\lambda}_t$$

$$\Sigma_{t|t} = (1-\rho_t)^{-1}\cdot(\Sigma_{t-1} + Q)$$

where $\rho_t = [\Sigma_{t-1} + Q][\Sigma_{t-1} + Q + R]^{-1}$