Variational Inference

$H(\varphi) \overset{\varphi \to \xi}{\triangleq} \frac{1}{2}\xi^\dagger \xi$

$$D_\varphi(d', d) := D_{KL}(Q \| P)$$

$$= \int \mathcal{D}\varphi \, Q(\varphi|d') \ln \frac{Q(\varphi|d')}{P(\varphi|d')}$$

let $Q(\varphi|d') = \mathcal{G}(\varphi - \bar{\varphi}, \Theta)$

$\bar{\varphi} \in \mathbb{R}^n$, $\Theta \in \mathbb{R}^{n \times n}$ with $n(n-1)/2$ d.o.f.

Wanted $\Theta = \Theta(\bar{\varphi})$ (n-dim parametrization)

remember Fisher Matrix

$$M(\varphi) := \left\langle \frac{\partial H(d|\varphi)}{\partial \varphi} \frac{\partial H(d|\varphi)^\dagger}{\partial \varphi} \right\rangle_{(d|\varphi)}$$

Post-Hamiltonian Hessian

$$\frac{\partial^2 H(\varphi|d)}{\partial \varphi \, \partial \varphi^\dagger} = \frac{\partial^2 H(d,\varphi)}{\partial \varphi \, \partial \varphi^\dagger} - \underbrace{\frac{\partial H(d)}{\partial \varphi \, \partial \varphi^\dagger}}_{=0}$$

$$= \frac{\partial^2 H(d,\varphi)}{\partial \varphi \, \partial \varphi^\dagger}$$

proposal

$$\Theta^{-1}(\theta) \approx \left\langle \frac{\partial^2 H(d,\varphi)}{\partial \varphi \, \partial \varphi^\dagger} \right\rangle_{(d|\varphi=\theta)}$$

$$= \left\langle \frac{\partial^2 H(\varphi)}{\partial \varphi \, \partial \varphi^\dagger} + \frac{\partial^2 H(d|\varphi)}{\partial \varphi \, \partial \varphi^\dagger} \right\rangle_{(d|\varphi=\theta)}$$

$[\ldots, M(\theta)]^{-1}$

$$\Theta^{-1}(\theta) \approx \ldots = \left\langle \mathbb{1} - \frac{\partial^2 \ln P(d|\xi)}{\partial \xi\, \partial \xi^+} \right\rangle_{P(d|\xi=\theta)}$$

$$= \mathbb{1} - \left\langle \frac{1}{P(d|\xi)} \frac{\partial^2 P(d|\xi)}{\partial \xi\, \partial \xi^+} \right\rangle$$

$$+ \left\langle \frac{1}{P^2(d|\xi)} \frac{\partial P(d|\xi)}{\partial \xi} \frac{\partial P(d|\xi)}{\partial \xi}^+ \right\rangle_{(d|\xi=\theta)}$$

$$= \mathbb{1} - \int dd \, \cancel{\frac{P(d|\xi)}{P(d|\xi)}} \frac{\partial^2 P(d|\xi)}{\partial \xi\, \partial \xi^+}$$

$$+ \left\langle \frac{\partial \ln P(d|\xi)}{\partial \xi} \frac{\partial \ln P(d|\xi)}{\partial \xi} \right\rangle_{(d|\xi=\theta)}$$

$$= \mathbb{1} - \underbrace{\frac{\partial^2}{\partial \xi \partial \xi^+} \underbrace{\int dd\, P(d|\xi)}_{1}}_{0} + \underbrace{\left\langle \frac{\partial H(d|\xi)}{\partial \xi} \frac{\partial H(d|\xi)}{\partial \xi}^+ \right\rangle_{(d|\theta)}}_{M(\theta)}$$

$$\Rightarrow \boxed{\Theta^{-1}(\theta) \approx \mathbb{1} + M(\theta)} > 0$$

since ≥ 0

(4)

minimize $f(\theta) = D_{KL}\left(\mathcal{G}(\xi - \theta, \Theta(\theta)) \,\|\, P(\xi|d)\right)$ w.r.t θ

trick: $\int d\xi\, \mathcal{G}(\xi - \theta, \Theta(\theta))\, f(\xi)$

$\approx \frac{1}{N} \sum_{i=1}^{N} f(\xi_i)$ with $\xi_i \hookleftarrow \mathcal{G}(\xi_i - \theta, \Theta(\theta))$, i.e.,

$\Rightarrow$ Metric Gaussian Variational Inference (MGVI)

Why is $\Theta(\theta) = (\mathbb{1} + M(\theta))^{-1}$ reasonable approx?

Linear, Gaussian case: $d = R\xi + n$

$$\mathcal{P}(\xi, n) = \mathcal{G}(\xi, \mathbb{1})\,\mathcal{G}(n, N)$$

$$\mathcal{P}(s|d) = \mathcal{G}(s-m, D), \quad D = (\mathbb{1} + R^{+}N^{-1}R)^{-1}$$

$$M(\theta) = \left\langle \frac{\partial H(d|\xi)}{\partial \xi} \frac{\partial H(d|\xi)}{\partial \xi}^{+} \right\rangle_{(d|\xi)}$$

$$H(d|\xi) \hat{=} \frac{1}{2}(d-R\xi)^{+}N^{-1}(d-R\xi)$$

$$\frac{\partial H(d|\xi)}{\partial \xi} = R^{+}N^{-1}(d-R\xi)$$

$$M(\xi) = \int dd\, \mathcal{G}(\underbrace{d-R\xi}_{n}, N) R^{+}N^{-1}(\underbrace{d-R\xi}_{n})(\underbrace{d-R\xi}_{n})^{+}N^{-1}R$$

$$= \int dn\, \mathcal{G}(n, N) R^{+}N^{-1}\, n n^{+} N^{-1}R$$

$$= R^{+}N^{-1}\underbrace{\langle n n^{+}\rangle_{\mathcal{G}(n,N)}}_{N} N^{-1}R = R^{+}N^{-1}R = M$$

$$\Rightarrow \Theta = (\mathbb{1} + M)^{-1} = D \text{ is exact}$$

$$\Rightarrow$$

nonlinear case:

directions constrained by data $\to$ $(\mathbb{1} + M(\theta))^{-1}$ is approximated (based on linearization)

directions not constrained by data $(\mathbb{1} + M(\theta))^{-1} \approx \mathbb{1}$ is correct prior covariance.

MGVI loss:

$$\mathcal{L}(\theta) = \int d\xi \; \mathcal{G}(\xi-\theta, \Theta) \ln \frac{\mathcal{G}(\xi-\theta, \Theta(\theta)) \, P(d)}{P(\xi|d) \, P(d)}$$

$$= \int d\xi \; \mathcal{G}(\xi-\theta, \Theta(\theta)) \ln \frac{\mathcal{G}(\xi-\theta, \Theta(\theta))}{P(d,\xi)} - \ln P(d)$$

$$\hat{=} \; S_B(\Theta(\theta)) + \langle H(d,\xi) \rangle_{\mathcal{G}(\xi-\theta, \Theta(\theta))}$$

drop to compute

$$\hat{\approx} \; \langle H(d,\xi) \rangle_{\mathcal{G}(\xi-\theta, \Theta(\theta))}$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} H(d, \xi_i) \qquad \xi_i \hookleftarrow \mathcal{G}(\xi_i - \theta, \Theta)$$

Use quasi Newton scheme (based on CG) with $\Theta(\theta)$ as proxy for second derivative,

Posterior Sampling from a Gaussian

$\mathcal{P}(s|d) = \mathcal{G}(s-m, D)$ how to get posterior samples?

$m = D R^{\dagger} N^{-1} d \qquad D = (S^{-1} + R^{\dagger} N^{-1} R)^{-1}$

observation: D is independent of d and thus from m

$\Rightarrow$ uncertainty variance can be transferred

$\Rightarrow$ generate synthetic signal & data s', d'

$\Rightarrow$ transfer synthetic reconstruction error $m' - s'$

$\Rightarrow$ get posterior sample for real data d

$s' \hookleftarrow \mathcal{G}(s', S), \quad n' \hookleftarrow \mathcal{G}(n', N), \quad d' = R s' + n'$

$m' = D R N^{-1} d', \qquad \Delta' = s' - m'$ (missed synthetic signal)

use CG here

sample $z_{\pm} = m \pm \Delta' = m \pm (s' - m')$

antithetic samples, thanks to symmetry

Claim $\mathcal{P}(z_{\pm}|d) = \mathcal{G}(s-m, D)$

<u>proof of claim</u> (2)

$$P(z_\pm|d) \stackrel{?}{=} \mathcal{G}(z_\pm - m, D) \Leftrightarrow P(\Delta'|d) \stackrel{?}{=} \mathcal{G}(\Delta', D)$$

$$z_\pm = m \pm \underbrace{(s' - m')}_{\Delta'}$$

$$D = (S^{-1} + \underbrace{R^\dagger N^{-1} R}_{M > 0})^{-1}$$

<u>moments of Δ':</u>

$$\boxed{\begin{aligned} \Delta' &= s' - DR^\dagger N^{-1}(Rs' + n') \\ &= s' - DMs' + DR^\dagger N^{-1} n = (1 - DM)s' + DR^\dagger N^{-1} n \\ &1 - DM > 0 \text{ since} \\ &1 - DM > 1 - D(S^{-1} + M) = 1 - DD^{-1} = 0 \end{aligned}}$$

$$\langle \Delta' \rangle_{(s', n')} = \langle (1 - DM)s' + DR^\dagger N^{-1} n' \rangle_{(s', n')} = (1 - DM)\langle s' \rangle_{(s')} + DR^\dagger N^{-1} \langle n' \rangle_{(n')} = 0$$

$$\langle \Delta' \Delta'^\dagger \rangle_{(s', n')} = (1 - DM) \underbrace{\langle s' s'^\dagger \rangle_{(s')}}_{S'} (1 - DM)^\dagger + 0 + DR^\dagger N^{-1} \underbrace{\langle n' n'^\dagger \rangle_{(n')}}_{N} N^{-1} R D + 0$$

since $\langle s' n'^\dagger \rangle_{(s', n')} = 0$

$$= (1 - DM) S' (1 - DM)^\dagger + D \underbrace{R^\dagger \underbrace{N^{-1} N N^{-1}}_{N^{-1}} R}_{M} D$$

$$\cancel{= S' - DMS' - S'MD + DMD} \quad \cancel{M = D^{-1} - S^{-1}}$$

$$\cancel{= S' - D[D^{-1} - S^{-1}]S' - S'(D^{-1} - S^{-1})D + D(D^{-1} - S^{-1})D}$$

$$\cancel{= S' - S' + D - S' + D + D - DS^{-1}D}$$

$$\cancel{= 3D - S' - DS^{-1}D}$$

$$\langle \Delta' \Delta^{\dagger\prime} \rangle_{(s', n')}$$

$$= S' - DMS - S'MD + DMSMD + DMD$$

$$DM = D\left(D^{-1} - S^{-1}\right) = \mathbb{1} - DS^{-1}$$

$$DMS = S' - D$$

$$\langle \Delta' \Delta'^{\dagger} \rangle_{(s', n')} = S' - 2(S' - D) + (S - D)(\mathbb{1} - S^{-1}D) + (\mathbb{1} - DS^{-1})D$$

$$= -S' + 2D + S' - D \underbrace{- D + DS^{-1}D + D - DS^{-1}D}_{0}$$

$$= D$$

n', s' have Gaussian statistics $\Rightarrow \Delta'$ as a linear combination of those is Gaussian as well

$$\Rightarrow P(\Delta' | d) = \mathcal{G}(\Delta' - \underbrace{\langle \Delta' \rangle}_{0}, \langle \Delta' \Delta'^{\dagger} \rangle)$$

$$= \mathcal{G}(\Delta', D)$$

$$z_{\pm} = m \pm \Delta$$

$$\Rightarrow P(z_{\pm} | d) = \mathcal{G}(z_{\pm} - m | D) \quad \text{qed}$$

Note 1:

Usage of CG in $m' = D \, \underbrace{R N^{-1} d'}$ - solving can leave some directions too small if terminated early $\Rightarrow$ $\Delta' = s' - m'$ will be too large in those directions

$\Rightarrow$ $z_\pm = m \pm \Delta_\pm$ will have more variance in those direction $\Rightarrow$ conservative representation of uncertainties. ✓

Note 2:

In MGVI $\mathcal{P}(\xi|d) = \mathcal{G}(\xi - \theta, \Theta)$ is needed.

Identify: $S' = \mathbb{1}$, $m = \theta$, $\Theta = D = (S^{-1} + R^\dagger N^{-1} R)^{-1}$

$$\Theta = (1 + M(\theta))^{-1}$$

and thus $M(\theta) = R^\dagger N^{-1} R$

a) in case of VF: $d = R\xi + n$, $n \hookleftarrow \mathcal{G}(n, N)$, $\xi \hookleftarrow \mathcal{G}(\xi, \mathbb{1})$

$$M(\theta) = \left\langle \frac{\partial H(d|\xi)}{\partial \xi} \frac{\partial H(d|\xi)}{\partial \xi}^\dagger \right\rangle_{(d|\xi)}$$

$$= \left\langle R^\dagger N^{-1} \underbrace{(d - R\xi)}_{n} \underbrace{(d - R\xi)}_{n}^\dagger N^{-1} R \right\rangle_{(d|\xi=\theta)}$$

$$= R^\dagger N^{-1} N N^{-1} R = R^\dagger N^{-1} R \quad ✓$$

$$\Rightarrow \Theta_{\text{exact}} = D$$

b) general case with gaussian noise

$$d = f(\varphi) + n \quad , n \hookleftarrow \mathcal{G}(n, N)$$

$$\cancel{M(\theta) =}$$

$$H(d|\varphi) \hat{=} \frac{1}{2} (\underbrace{d - f(\varphi)}_{n})^{+} N^{-1} (d - f(\varphi))$$

$$\frac{\partial H(d|\varphi)}{\partial \varphi} = - \frac{\partial f(\varphi)}{\partial \varphi} N^{-1} n$$

$$M(\theta) = \left\langle \frac{\partial H(d|\varphi)}{\partial \varphi} \frac{\partial H(d|\varphi)}{\partial \varphi} \right\rangle_{(d|\varphi=\theta)} = \frac{\partial f(\varphi)}{\partial \varphi} \underbrace{N^{-1} \underbrace{\langle n n^{+} \rangle_{(n)}}_{N} N^{-1}}_{N} \frac{\partial f}{\partial \varphi}^{+} \Big|_{\varphi=\theta}$$

$$\Rightarrow M(\theta) = \underbrace{\frac{\partial f}{\partial \varphi}\Big|_{\varphi=\theta}}_{= R_\theta} \underbrace{N^{-1}}_{N^{-1}} \underbrace{\frac{\partial f}{\partial \varphi}\Big|_{\varphi=\theta}^{+}}_{R_\theta^{+}}$$

$\uparrow$

linearized Response

Exercise: calculate $M(\theta)$ for Poisson-data

$$P(d|\lambda) = \prod_x \frac{(\lambda^x)^{d^x} e^{-\lambda^x}}{d^x !}$$

Fisher Information for Poissonian data

$$H(d|\xi) = \lambda(\xi)^\dagger 1 - d^\dagger \ln \lambda(\xi) + \ln d!$$

$$\lambda(\xi) = R \exp(A_s \xi) \quad , \quad \frac{\partial \lambda^i(\xi)}{\partial \xi^k} = R^i_{\ x} (A_s)^x_{\ k}\, e^{A_s{}^x_{\ k} \xi^k}$$

$$\frac{\partial H(d|\xi)}{\partial \xi} = \frac{\partial \lambda^\dagger(\xi)}{\partial \xi}\left(1 - \frac{d}{\lambda}\right)$$

$$M(\xi) = \left\langle \frac{\partial H(d|\xi)}{\partial \xi} \frac{\partial H(d|\xi)}{\partial \xi}^\dagger \right\rangle_{(d|\xi)}$$

$$= \frac{\partial \lambda^i}{\partial \xi} \left\langle \frac{(\lambda^i - d^i)(\lambda^j - d^j)}{\lambda^i \lambda^j} \right\rangle_{(d|\xi)} \frac{\partial \lambda^j}{\partial \xi}$$

$$= \frac{\partial \lambda^i}{\partial \xi} \frac{\delta_{ij}\, \lambda^i}{\lambda^{i2}} \frac{\partial \lambda^j}{\partial \xi} = \frac{\partial \lambda}{\partial \xi} \widehat{\frac{1}{\lambda}} \frac{\partial \lambda^\dagger}{\partial \xi^\dagger}$$

$$= \frac{\partial \lambda(\xi)}{\partial \xi} \widehat{\lambda^{-1/2}}\, \widehat{\lambda^{-1/2}} \frac{\partial \lambda^\dagger}{\partial \xi^\dagger} \equiv \frac{\partial x(\xi)^\dagger}{\partial \xi} \frac{\partial x(\xi)}{\partial \xi^*} = \frac{\partial x^\dagger}{\partial \xi} \mathbb{1} \frac{\partial x}{\partial \xi}$$

Metric in x-space $\frac{ds^2}{dx\,dx}$

with $\frac{\partial x(\xi)}{\partial \xi} = \widehat{\lambda^{-1/2}} \frac{\partial \lambda}{\partial \xi} = \frac{1}{2} \frac{\partial \sqrt{\lambda(\xi)}}{\partial \xi}$

$$x(\xi) = \frac{1}{2}\sqrt{\lambda(\xi)}$$ coordinates that flatten the Fisher metric

more specifically: $M(\xi)$ is pullback of Euclidean metric

$$M(\xi) = \left(\frac{\partial x}{\partial \xi}\right)^\dagger \frac{\partial x}{\partial \xi}$$

Q: Can we as well flatten the MGVI metric $A^{-1} = \mathbb{1} + M$?

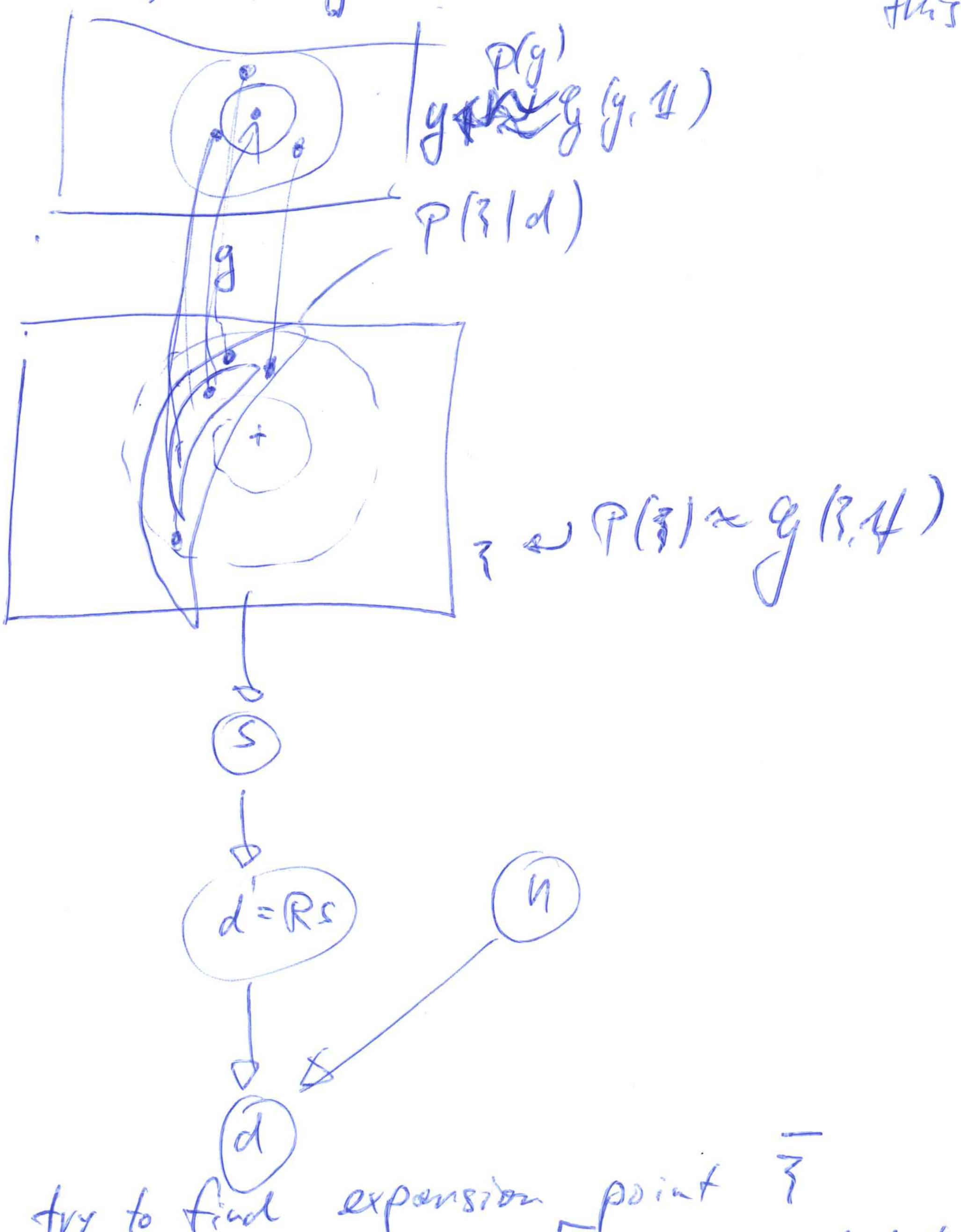
$$H \overset{?}{=} \left(\frac{\partial g(\xi)}{\partial \xi}\right)^\dagger \frac{\partial g(\xi)}{\partial \xi} \overset{!}{=} \mathbb{1} + \left(\frac{\partial x}{\partial \xi}\right)^\dagger \frac{\partial x}{\partial \xi} = M$$

in general: no, but approximatively

$$g(\xi;\bar\xi) = M_{\bar\xi}^{-1/2}\left[(\xi-\bar\xi) + \left(\frac{\partial x}{\partial \xi}\Big|_{\bar\xi}\right)^\dagger \left(x(\xi)-x(\bar\xi)\right)\right] + \mathcal{O}(\cancel{x}^2)$$

$$=: M_{\bar\xi}^{-1/2}\,\tilde g(\xi;\bar\xi)$$

Exercise: proof this



geoVI: try to find expansion point $\bar\xi$ for which $Q(\xi|\bar\xi) = \left[g(\cdot;\bar\xi)_* \, \mathcal{G}(\cdot|\mathbb{1})\right](\xi)$

push forward

minimizes $\mathcal{D}_{KL}(Q(\xi|\bar\xi)\,||\,P(\xi|d))$.

$$\hat{H}^{-1} \stackrel{?}{=} \left(\frac{\partial g(\zeta,\bar{\zeta})}{\partial \zeta}\right)^{+} \frac{\partial g(\zeta,\bar{\zeta})}{\partial \zeta}$$

$$\left.\frac{\partial g(\zeta,\bar{\zeta})}{\partial \zeta}\right|_{\bar{\zeta}} = M^{-1/2}\left[\mathbb{1} + \left(\left.\frac{\partial x}{\partial \zeta}\right|_{\bar{\zeta}}\right)^{+} \left.\frac{\partial x}{\partial \zeta}\right|_{\bar{\zeta}}\right]$$

$$= M^{-1/2}\left[\mathbb{1} + M\right]$$

$$\cancel{\hat{H}\Big|_{\bar{\zeta}} \stackrel{?}{=} \left[\mathbb{1} + M\right]^{+} \underbrace{M^{-1/2^{+}} M^{-1/2}}_{} \left[\mathbb{1} + M\right] = M^{-1} + 2\mathbb{1} + M}$$

$$\cancel{(M^{-1} + \mathbb{1})(\mathbb{1} + M) = \mathbb{1}}$$

$$\hat{H}^{-1} = (\mathbb{1} + M)\, M^{-1/2} M^{-1/2} (1 + M)$$

$$= (\mathbb{1} + M)(1 + M)^{-1}(1 + M)$$

$$= 1 + M = M \quad \checkmark$$