PS 3

(a)

$$D_{KL}(P\|Q) = \sum_{x\in X} P(x)\log\frac{P(x)}{Q(x)}$$

$$= -\sum_x P(x)\log\frac{Q(x)}{P(x)} \qquad \left(\approx \mathop{E}_{x\sim P}(\log Q/P)\right)$$

$$= -\mathop{E}_{x\sim P}\left[\log\frac{Q(x)}{P(x)}\right]$$

strict: log is concave, $-\log$ is convex

$$-D_{KL} = \mathop{E}_{x\sim P}[\,\cdot\,] \leq \log \mathop{E}_{x\sim P}\left[\frac{Q(x)}{P(x)}\right] \quad \text{①}$$

$$= \log \sum_x P(x)\frac{Q(x)}{P(x)}$$

$$= \log 1 = 0$$

$$\therefore D_{KL} \geq 0 \qquad \square$$

if $P = Q$ $\quad D_{KL} = \sum_x P(x)\log 1$

$$D_{KL} = 0$$

∵ log strict concave

$$D_{KL} = \mathop{E}_{x\sim P}\left[\left(\log\frac{P(x)}{Q(x)}\right)\right] = \log\left(\frac{P(x)}{Q(x)}\right)$$

∵ $E(x) = x$ as strict convex

need strict concave

$-\log$

$$-D_{KL} = \sum_x P(x)\log\frac{Q(x)}{P(x)}$$

$$= \mathop{E}_{x\sim P}\left[\log\frac{Q(x)}{P(x)}\right] = \log\left(\frac{Q(x)}{P(x)}\right)$$

① & antilog $\quad \mathop{E}_P\left[\frac{Q}{P}\right] = \sum P\frac{Q}{P} = 1 \longrightarrow \therefore \frac{Q(x)}{P(x)} = 1$ as $E(Q/P) = Q/P$

~~$\log E_P[\,\cdot\,] = \log 1 = 0$~~

$$\therefore P(x) = Q(x)$$

(b) $D_{KL}(P(x) \| Q(x)) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$

+

$D_{KL}(P(Y|X) \| Q(Y|X)) = \sum_x P(x) \left(\sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} \right)$

$= \sum_x P(x) \left(\log \frac{P(x)}{Q(x)} + \sum_y P(y|x) \log \frac{P(y|x)}{Q(y|x)} \right)$

~~$\sum_y \sum_x P(y|x) P(x)$~~

$D_{KL}(P(x,y) \| Q(x,y)) = \sum_{x,y} P(x,y) \log \frac{P(x,y)}{Q(x,y)}$

$= \sum_{x,y} P(x,y) \log \frac{P(x) P(y|x)}{Q(x) Q(y|x)}$

$= \sum_{x,y} P(x,y) \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(x,y) \log \frac{P(y|x)}{Q(y|x)}$

$= \sum_x \sum_y P(x|y) P(y) \log \frac{P(x)}{Q(x)} + \sum_{x,y} P(y|x) P(x) \log \frac{P(y|x)}{Q(y|x)}$

$\sum_x \log \frac{P(x)}{Q(x)} \sum_y P(x|y) P(y)$

$= \sum_x \left(\log \frac{P(x)}{Q(x)} \right) P(x) + \sum_{x,y} P(x) P(y|x) \log \frac{P(y|x)}{Q(y|x)}$

(c) $\hat{P}(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x^{(i)} = x\}$ uniform dist over training set

$P_\theta(x) \equiv P(x;\theta)$ $\quad \arg\min_\theta D_{KL}(\hat{P} \| P_\theta) = \arg\max_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)})$

$D_{KL}(\hat{P} \| P_\theta) = \sum_x \hat{P}(x) \log \frac{\hat{P}(x)}{P(x;\theta)}$

$= \sum_x \hat{P}(x) \log \hat{P}(x) - \sum_x \hat{P} \log P(x,\theta)$

argmin wrt θ : $\arg\min_\theta - \sum_x \hat{P} \log P_\theta(x) = \arg\max_\theta \sum \hat{P} \log P(x)$

$= \arg\max_\theta \sum_x \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{x^{(i)} = x\} \log_2 P_\theta(x)$

$\frac{1}{n} \sum_x \sum_{i=1}^{n} \mathbb{1}\{x^{(i)} = x\} \log_2 P_\theta(x)$

$\frac{1}{n} \sum_{i=1}^{n} \sum_x \mathbb{1}\{x^{(i)} = x\} P_\theta(x) = \arg\max_\theta \sum_{i=1}^{n} \log P_\theta(x^{(i)})$

MLE θ = min KL from $\hat{P}$

Q2 (b) 24 bits to 16 color

(8×3) ∴ need $\log_2(16)$ bits = 4

$\therefore \frac{24}{4} = 6$ compression

Q3 (a) $l_{semisup}(\theta^{(t+1)}) = \sum_{i=1}^{n} \log \sum_z Q \frac{P(x^{(i)}, z^{(i)}; \theta)^{(t+1)}}{Q} + \alpha \log p(\theta^{(t+1)})$ $\quad E_Q\left(\frac{P}{Q}\right)$

Jensen $\geq \sum_{i=1}^{n} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)}; \theta^{(t+1)})}{Q_i(z^{(i)})} + \alpha \log p(\theta^{(t+1)})$

M Step $\geq$ ——— " ——— $; \theta^{(t)}$ $\quad \alpha \log p(\theta^{t})$

$\therefore Q_i(z^{(i)}) = P(z^{(i)} \mid x^{(i)}; \theta)$ & $\sum Q_i(z^{(i)}) = 1$

$x \cap z = x|z \cdot z = z|x \cdot x$ $\quad \therefore \frac{P(x \cap z)}{P(z|x)} = P(x)$

$= \sum_{i=1}^{n} \log \frac{P(x, z; \theta)}{P(z|x; \theta)} + \alpha l_{sup}$

$= \sum_{i=1}^{n} \log P(x^{(i)}; \theta^{(t)}) + \alpha l_{sup}(\theta^{t})$

$= l_{semisup}(\theta^{(t)})$

(b) latent is still $z^{(i)}$, $i \in \{1, \dots, n\}$
∴ E step like ~~out~~ unsupervised

$$\therefore w_j^{(i)} = Q_i(z^{(i)} = j) = \frac{p(x^{(i)} | z^{(i)} = j; \cdot)\, p(z^{(i)} = j)}{\sum_{l=1}^{k} p(x^{(i)} | z^{(i)} = l_i)\, p(z^{(i)} = l)}$$

$$C_{j,l} = \frac{1}{(2\pi)^{q/2} |\Sigma_{j,l}|^{1/2}}$$

$$\therefore = C_j \frac{\exp\{-1/2 (x^{(i)} - \mu_j)^T \Sigma_j^{-1} (x^{(i)} - \mu_j)\} \phi_j}{\sum_{l=1}^{k} C_l \exp\{\frac{-1}{2}(x^{(i)} - \mu_l)^T \Sigma_l^{-1} (x^{(i)} - \mu_l)\} \phi_l}$$

(c) $ELBO_i^{(t)}(\theta) = \underset{z^{(i)} \sim Q_i(z^{(i)})}{\mathbb{E}} \log \left[\frac{P(x^i | z^i; \theta^t) P(z^i; \theta^t)}{Q_i(z^i = j)} \right]$ unsup
$+ \alpha \left[\log P(\tilde{x}_j^{(i)} | z^{(i)}; \theta^t) \frac{P(z^{(i)}; \theta^t)}{} \right]$ sup
i data pt.

$\sum_{z^{(i)}} Q_i(z^{(i)}) \cdot \log[-] + \alpha [\cdot]$ ← $l_{sup}(\theta^{(t)})$

$\tilde{w}_j^{(i)} = \mathbb{1}\{\tilde{z}^{(i)} = j\}$ & $w_j^{(i)} = Q_i(z^{(i)} = j)$
$Q_i(z^{(i)}) = P(z^{(i)} | x^{(i)}; \theta^{(t)})$

GMM: $\theta^{(t)} = (\mu_j^{(t)}, \sigma_j^{(t)}, \phi_j^{(t)})$

$$\theta^{(t+1)} = \arg\max_{\theta} l_{semisup}(\theta^{(t)})$$

$$\nabla_\theta l_{ss} = \nabla_\theta \sum_{i=1}^{n} ELBO_i^{(t)}$$

$$= \nabla_\theta \left(\sum_{i=1}^{n} \underset{z^{(i)} \sim Q_i(z^{(i)})}{\mathbb{E}} \log[-] + \alpha \sum_{i=1}^{\tilde{n}} l_{sup}^i(\theta^{(t)}) \right)$$ ran var all θ

$\frac{\partial \nabla l_{ss}}{\partial \mu_j^{(t)}, \Sigma_j^{(t)}, \phi_j^{(t)}}$ in order $\nabla = 0$ wrt all param

l_{ss}: $Q(\cdot) [\log(p(x|z)) + (\log \frac{P(z)}{Q(z=j)}) + \alpha(x|z) + p(x)$

$\arg\min l_{ss} = \sum Q(\cdot) \log_a(p(x|z)) + \sum \alpha(x|z)$

$$= \sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q_i(z^{(i)})} \left[\nabla_{\mu_j} \log p(x^{(i)} | z^{(i)}; \theta^{(t)}) \right] + \alpha \sum_{i=1}^{\tilde{n}} \left[\nabla_{\mu_j} \log p(\tilde{x}^{(i)} | \tilde{z}^{(i)}; \theta^{(t)}) \right]$$

- wrt μ_j : $\nabla \text{const} = 0$ of gaussian

$$\sum_{i=1}^{n} \mathbb{E}_{z^{(i)} \sim Q_i(z^{(i)})} \left[\nabla_{\mu_j} \left(-\tfrac{1}{2} (x^{(i)} - \mu_j)^T \Sigma^{-1} (x^{(i)} - \mu_j) \right) \right] + \dots (\tilde{x}^{(i)} - \mu_j)^T \Sigma^{-1} (\tilde{x}^{(i)} - \mu_j)$$

$Q_i(z^{(i)}) = w_j^{(i)}$

$$\sum_{i=1}^{n} w_j^{(i)} \Sigma^{-1} (x^{(i)} - \mu_j) + \alpha \sum_{i=1}^{\tilde{n}} \Sigma^{-1} (\tilde{x}^{(i)} - \mu_j) = 0$$

$$\Leftrightarrow \sum_{i=1}^{n} w_j^{(i)} \Sigma^{-1} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \Sigma^{-1} (\tilde{x}^{(i)})$$

$$\therefore \mu_j = \frac{\sum_{i=1}^{n} w_j^{(i)} x^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} \tilde{x}^{(i)}}{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

- Σ_j : using unsup GMM formula for Σ_j gives analogous Σ_j for unsup

$$\Sigma_j = \frac{\sum_{i=1}^{n} w_j^{(i)} (x^{(i)} - \mu_j)^T (x^{(i)} - \mu_j) + \alpha \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)} (\tilde{x}^{(i)} - \mu_j)^T (\tilde{x}^{(i)} - \mu_j)}{\sum_{i=1}^{n} w_j^{(i)} + \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}$$

③ lagrange $\sum_{j=1}^{m} \phi_j = 1$

analogy gives $\phi_j = \frac{\sum_{i=1}^{n} w_j^{(i)} + \alpha \sum_{i=1}^{\tilde{n}} \tilde{w}_j^{(i)}}{n + \alpha \tilde{n}}$

$\tilde{w}_j^{(i)} = \mathbb{1}\{\tilde{z}^{(i)} = j\}$ we optimize $\log p(\tilde{z}^{(i)}; \theta^{(t)})$

completeness : $\mathcal{L}(\phi, \lambda) = C + \sum_{i=1}^{n} \sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \alpha \sum_{i=1}^{\tilde{n}} \sum_{j=1}^{k} \tilde{w}_j^{(i)} \times \log \phi_j + \lambda \left(\sum_{j=1}^{k} \phi_j - 1 \right)$

$\nabla_{\phi_j, \lambda} = 0$

(4) (i) SS: faster (25 iter) compared to US ~150

(ii) SS: stable US: varies with initialization

(iii) US: error - failed single high variance gauss

SS: almost exact as $\tilde{x}$ higher quality underlying

2018

Q1 (a) $\frac{\partial \ell}{\partial w_{ij}^{[1]}} = \frac{\partial \ell}{\partial o} \cdot \frac{\partial o}{\partial h_2} \cdot \frac{\partial h_2}{\partial w_{12}^{[1]}}$ (1) (2) (3)

$i=1, j=2$

$\sigma(z) = \frac{1}{1+e^{-z}}$ $\qquad z_o$

$o^{(i)} = \sigma\left(\sum_{j=1}^{3} w_j^{[2]} h_j^{(i)} + w_0^{[2]}\right)$

z_{h_2}

$h_j^{(i)} = \sigma\left(\sum_{k=1}^{2} w_{kj}^{[1]} x_k^{(i)} + w_{0,j}^{[1]}\right)$

(1) $\frac{\partial \ell^{(i)}}{\partial o^{(i)}} = 2(o^{(i)} - y^{(i)})$

(2) from $\sigma(z)$,

$\frac{\partial o^{(i)}}{\partial z_o} = o^{(i)}(1-o^{(i)})$

$\frac{\partial o^{(i)}}{\partial h_2^{(i)}} = \frac{\partial o^{(i)}}{\partial z_o} \cdot \frac{\partial z_o}{\partial h_2^{(i)}} = o^{(i)}(1-o^{(i)}) w_2^{[2]}$

(3) $\frac{\partial h_2^{(i)}}{\partial z_{h_2}} = h_2^{(i)}(1-h_2^{(i)})$ sigmoid

$\therefore \frac{\partial h_2^{(i)}}{\partial w_{12}^{[1]}} = \frac{\partial h_2^{(i)}}{\partial z_{h_2}} \cdot \frac{\partial z_{h_2}}{\partial w_{12}^{[1]}}$

$= h_2^{(i)}(1-h_2^{(i)}) \cdot x_1^{(i)}$

$\therefore$ Combining $2(o^{(i)} - y^{(i)}) \cdot o^{(i)}(1-o^{(i)}) w_2^{[2]} h_2^{(i)}(1-h_2^{(i)}) x_1^{(i)}$

$w_{12}^{[1]} := w_{12}^{[1]} - \alpha \frac{1}{m} \sum_{i=1}^{m} \frac{\partial \ell^{(i)}}{\partial w_{12}^{[1]}}$

(b) yes

- hidden layer neurons act as threshold & partition the input space
- weigh affects boundary $\quad y=0.5 \quad x=0.5 \quad x+y=1$
- 3 neurons can act as 3 sides of the Δ to partition

(c) No: if linear, essentially output thresholds on linear combination giving boundary like perceptron. data is not linearly separable

0.5

(b) $w_{11}x_1 + w_{21}x_2 + b_1 = z_1$ → $0 \cdot x_1 + x_2 - 0.5 = z_1$

$w_{12}x_1 + w_{22}x_2 + b_2 = z_2$ $x_1 + 0 \cdot x - 0.5 = z_2$

$w_{13}x_1 + w_{23}x_2 + b_3 = z_3$ $x_1 + x_2 - 4 = z_3$

0.5 0.5 0

$h_j = f(z_j)$

$o = f(w_{h_1}h_1 + w_{h_2}h_2 + w_{h_3}h_3 + b_0)$

1 1 −0.5

1 1 0 0.5 ①

1 −1 1 0.5

$h_1 = 1$ $h_2 = 1$ $h_3 = -1$ $b_0 = 0.5$

$b_0 = -1.5$

$w_{h_3} = -1$

$w_{h_1} = w_{h_2} = 1$

$w_{11} = 0$ $b_1 = -0.5$

$w_{21} = 1$

similar for w_{12}, w_{22}, b_2

$w_{13} = w_{23} = 1$ $b_3 = -4$

eg pt. 4, 4

$z_2 = z_1 = 3.5$ $h_1 = 1 = h_2$ $z_3 = 4$ $h_3 = 1$

$0 = f(1 + 1 + (-1)(1) - 0.5)$

$= f(0.5)$ assuming

$= 1$

pt 1 1

$z_1 = z_2 = 0.5$ $h_1 = h_2 = 1$ $z_3 = -2$ $h_3 = 0$

$0 = f(1 + 1 + (-1)(0) - 0.5)$

pt .0.5, 0.5

0, 0, 1, 1 −4 0

$(1 + 1 + 0 - 1.5) = f(0.5) \therefore 1$ ✓

pt 2, 2

z_4 −1.5 1.5 1 1 0 1

$1 + 1 + (-1)(1) - 0.5 = f(0.5) = 1$ ✗

∴ bias bigger

pt, 0, 0

0, 0, 0

$0 + 0 + 0 - \quad f(-n) = 0$ ✓

BRUNNEN

(2) $D_{KL}(p\|q) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$

$= \underset{x\sim p}{E}(\log p(x)) - \underset{x\sim p}{E}[\log(q(x))]$

(a) Show $\underset{y\sim p(y;\theta)}{E}[\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}] = 0$

score of $p(y;\theta) = \nabla_\theta \log p(y;\theta)$

$= \frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}$

$\therefore \underset{y\sim p(y;\theta)}{E}\left[\frac{\nabla_\theta p(y;\theta)}{p(y;\theta)}\right]$

$= \int_{-\infty}^{\infty} p(y;\theta) \cdot \frac{\nabla_\theta p(y;\theta)}{p(y;\theta)} dy$

$= \int_{-\infty}^{\infty} \nabla_\theta p(y;\theta) dy$

Applying Leibniz

$= \nabla_\theta \int_{-\infty}^{\infty} p(y;\theta) dy = \nabla_\theta 1 = 0$

(b) Fisher $I(\theta) = \underset{y\sim p(y;\theta)}{Cov}[\nabla_{\theta'} \log p(y;\theta')|_{\theta'=\theta}]$ Cov of score

Show $= \underset{y\sim p(y;\theta)}{E}[\nabla_\theta \log p(y;\theta) \nabla_\theta \log p(y;\theta)^T]$ — (1)

$Cov = \frac{1}{n} XX^T \simeq E[XX^T] = E[(x - E(x))(x - E(x))^T]$

$\therefore$ Using (1) $\rightarrow \int_{-\infty}^{\infty} \nabla_\theta \log p (\nabla_\theta \log p) \cdot p(y;\theta) dy$

in discrete, $cov(\nabla \log p) = \frac{1}{n} \nabla \log p \nabla \log p^T$

© Show $\mathbb{E}_{y\sim p(y;\theta)}[-\nabla_\theta^2 \log p(y;\theta)] = \mathcal{I}(\theta)$ $\mathbb{E}[-\text{Hessian}]$

$$\nabla_\theta \log p = \frac{\nabla_\theta p}{p}$$

$$\nabla_\theta(\cdot) = \nabla_\theta\left(\frac{\nabla_\theta p}{p}\right) = (\nabla_\theta p)\left(\frac{-\nabla_\theta p}{p^2}\right) + \frac{1}{p}\nabla_\theta^2 p$$

$$-\nabla_\theta(\cdot) = \frac{(\nabla_\theta p)^2 - p\nabla_\theta^2 p}{p^2}$$

$$-\nabla_\theta^2 p \frac{(\nabla_\theta p)^2}{p^2} + \frac{p\nabla_\theta^2 p}{p^2}$$

$$I(\theta) = \mathbb{E}\left[\frac{\nabla p}{p}\frac{\nabla p}{p}\right] = \int_{-\infty}^{\infty} \frac{(\nabla p)^2}{p^2} p\, dy$$

$$\mathbb{E}(-\nabla_\theta(\cdot)) = \int p\frac{(\nabla_\theta p)^2}{p^2} dy - \int \nabla_\theta^2 p\, dy$$

$$-\nabla_\theta^2 \int_0 p\, dy$$

$$\therefore \mathbb{E}(-\nabla_\theta(\cdot)) = I(\theta)$$

(d) $\log p(y;\tilde\theta) \approx \log p(y;\theta) + (\hat\theta-\theta)^T \nabla_{\theta'} \log p(y;\theta')\big|_{\theta'=\theta}$ (Taylor around θ)

$$+\frac{1}{2}(\hat\theta-\theta)^T(\nabla_{\theta'}^2 \log p(y;\theta')\big|_{\theta'=\theta})(\hat\theta-\theta)$$

$\tilde\theta = \theta + d$

$$= \log p(y;\theta) + d^T \nabla_{\theta'} \log p\big|_{\theta'=\theta} + \frac{1}{2}(d)^T(\nabla_\theta^2 \log p\big|_{\theta'=\theta})(d)$$

$$\mathbb{E}_{y\sim p(y;\theta)}(\log p(\tilde\theta)) = \mathbb{E}[\log p] + d^T \underbrace{\mathbb{E}[\nabla_\theta \log p]}_{0} + \frac{1}{2}d^T \mathbb{E}[\nabla_\theta^2 \log p](d)$$

$$\therefore \mathbb{E}[\log p]_{\tilde\theta} = \mathbb{E}[\log p]_{\theta} + \frac{1}{2}d^T I(\theta) d$$

$$D_{KL}(p_\theta \| p_{\tilde\theta}) = \mathbb{E}_{x\sim p(y;\theta)}[\log p(y;\theta)] - \mathbb{E}_{y\sim p(y;\theta)}[\log p(y;\tilde\theta)]$$

$$\approx \frac{1}{2}d^T I(\theta) d$$

(c) $\ell(\theta+d) \approx \ell(\theta) + d^T \nabla_{\theta'} \ell(\theta')|_{\theta'=\theta}$

$= \log p(y;\theta) + d^T \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)}$

$D_{KL}(p_\theta \| p_{\theta'}) \approx \frac{1}{2} d^T I(\theta) d$

$L(d,\lambda) = \ell(\theta+d) - \lambda[D_{KL}(p_\theta \| p_{\theta'}) - c]$

$\approx \log p(y;\theta) + d^T \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)} - \lambda\left[\frac{1}{2} d^T I(\theta) d - c\right]$

$\nabla_d L(d,\lambda) = 0 + \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)} - \lambda I(\theta) d \overset{!}{=} 0$

$d = \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)} \cdot \frac{1}{\lambda I(\theta)} = \frac{1}{\lambda}(Z)$

$\nabla_\lambda L(d,\lambda) \overset{!}{=} 0 = \frac{1}{2} d^T I(\theta) d - c$

$2c = d^T I(\theta) d$

$= \frac{1}{\lambda} Z^T I(\theta) \frac{1}{\lambda} Z$

$\lambda = \pm \sqrt{Z^T I(\theta) Z / 2c}$

$d^* = \frac{1}{\lambda} Z = \frac{Z}{\sqrt{Z^T I(\theta) Z}} \sqrt{2c}$ with $Z = \frac{\nabla_{\theta'} p(y;\theta')|_{\theta'=\theta}}{p(y;\theta)} \frac{1}{I(\theta)}$

(d) Newton: $\theta := \theta - H^{-1} \nabla_\theta \ell(\theta)$

Natural grad: $I(\theta) = \mathbb{E}_{y \sim p(y;\theta)}[-\nabla^2_{\theta'} \log p(y;\theta)|_{\theta'=\theta}]$

$-\nabla^2_\theta \ell(\theta) = H$

$\theta := \theta + \tilde{d} = \theta + \frac{1}{\lambda} I(\theta)^{-1} \nabla_\theta \ell(\theta)$