Specialized MCMC

Adaptive MCMC

Ex: Nueral Networks

data: $\{(y_i, x_i)\}_{i}^{n}$ $\quad x_i \in \mathbb{R}^p$

model: $y_i = f_\theta(x_i) + \varepsilon_i \quad i = 1, \dots, n$

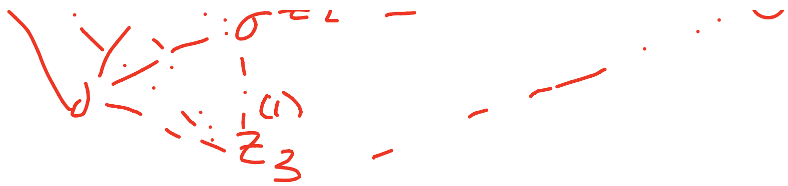
$$\varepsilon_i \sim N(0, \sigma^2)$$

where $f_\theta(x) = \sum_{k=1}^{K} b_k \, s(a_{0k} + \langle a_k, x \rangle)$

$$\theta = \left((a_{0,1}, a_1, b_1), \dots, (a_{0,K}, a_K, b_K) \right) \in \mathbb{R}^{K(p+1)}$$

$s(\cdot)$ activation function

$$z_1^{(1)} = s(a_{01} + \langle a_1, x \rangle)$$



Class of nonlinear regression functions

<u>Bayesian</u>: Prior on Θ

$$p(\theta) = \prod_{k=1}^{K} p\left(a_{0k}, a_k, b_k\right)$$

$$= \prod_{k=1}^{K} \underbrace{p(a_{0k})}_{N(0,\sigma_{0k}^2)} \; \underbrace{p(a_k)}_{N(0,\sigma_a^2 I_{p+1})} \; \underbrace{p(b_k)}_{N(0,\sigma_b^2)}$$

Therefore

$$\pi(\theta \mid y_{1:n}) \propto p(\theta)\left(\frac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i - f_\theta(x_i)\right)^2\right]$$

<u>Gibbs</u>: Update one component at a time.

$$b_k \mid (\theta \setminus b_k) \propto p(b_k) \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2 \right\}$$

This is intractable so we do a Metropolis to update θ.

Choose a proposal: e.g. a R.W. metropolis.

<u>Alg</u>: Given $\theta^{(t)} \in \mathbb{R}^{K(p+1)}$

Randomly choose $j \in [1, \ldots, K(p+1)]$

Sample $\xi = N(\theta_j^{(t)}, v_j^2)$

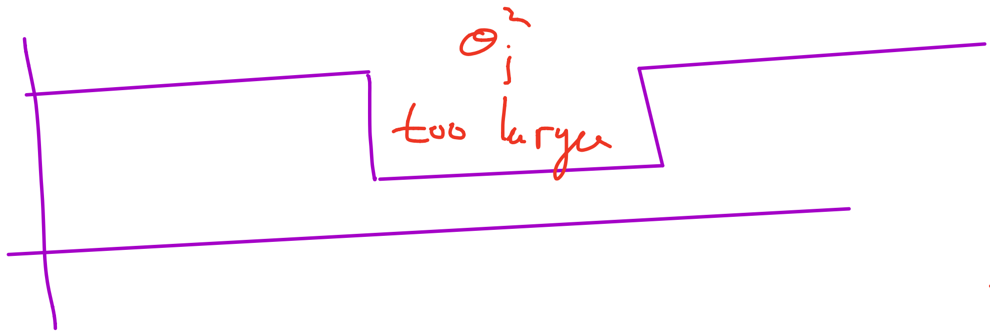
Set. $$\theta_j^{(t+1)} = \begin{cases} \xi & \text{with prob } A_t^{*} \\ \theta_j^{(t)} & \text{---- } 1 - A_t \end{cases}$$

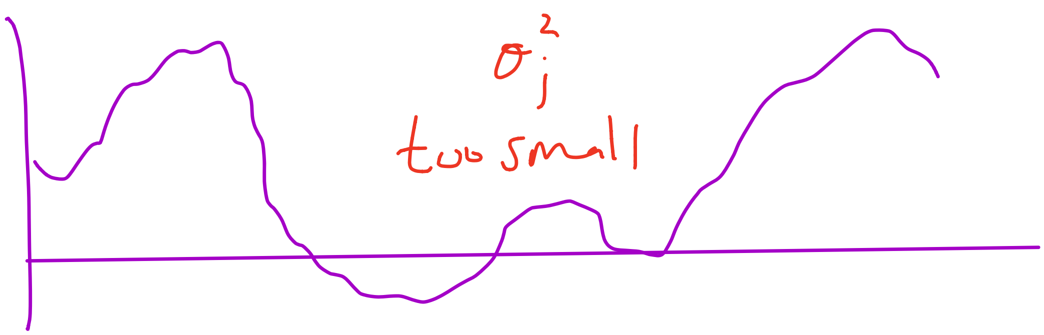
set $\theta_\ell^{(t+1)} = \theta_\ell^{(t)} \quad \ell \neq j$

* when $$A_t = \min\left\{ 1, \frac{\pi(\xi, \mid \theta^{(t)}, y)}{\pi(\theta_j^{(t)} \mid y, \theta^{(t)})} \right\}$$

Rmk: $v := (v_1^2, \dots, v_j^2, \dots, v_{K(p+1)}^2)$

is the vector of step sizes.





Q: Can we adaptively update step sizes so that we achieve the sweet spot of acceptance?

$$\theta^{(0)} \sim \pi(\cdot \mid y_{1:n}) \qquad \theta^{(1)} \mid \theta^{(0)} \sim P_v(\theta^{(0)}, \cdot)$$

$$\psi(v) := \mathbb{E}\left[\|\theta^{(1)} - \theta^{(0)}\|_2^2\right]$$

Goal: $V_{opt} = \max_{v} \psi(v)$

explore most area in 1 step.

Update the step size

$$v^{(t+1)} = v^{(t)} + \gamma^{(t+1)} \nabla \psi(v^{(t)}).$$

Ex: Suppose $\pi : (0, \infty)$ then the kernel is given by

$$P_\sigma(x, A) = \int_A \min(\ast)\, \phi_\sigma(x,y)\,dy + (1 - A_t)\, 1_A(x)$$

$$\psi(\sigma) := \mathbb{E}\left[\|X_1 - X_0\|_2^2 \right]$$

$$= \int \pi(x) \int \min(\ast)(y-x^2)\, \phi_\sigma(x)\, dy\, dx$$

$$\frac{d}{d\sigma^2} \psi(\sigma^2) = \iint \min(\ast)\,(y-x)^2 \left(\frac{1}{2\sigma^4}(y-x)^2 - \frac{1}{2}\,\frac{1}{\sigma^2}\right) \times \phi_\sigma(x,y)\,\pi(x)\,dy\,dx$$

$$= \frac{1}{2\sigma^2}\,\mathbb{E}\left[(Y-X)^2\left(\frac{(Y-X)^2}{\sigma}-1\right)\min\left(1\ \frac{\pi(Y)}{\pi(X)}\right)\right]$$

Therefore we can plugin (ξ, X) into this expression and use the N.R. update.

$$V_j^{(t+1)} = V_j^{(t)} + \frac{\gamma^{(t+1)}}{2\sigma^2}\left\{(\xi-\theta_j^{(t)})^2\right.$$

$$\times \left[\left(\xi - \theta_j^{(t)} \right)^2 - 1 \right] A_t \}$$

Normally require

$$\gamma^{(t)} > 0 \qquad \sum \gamma^{(t)} = +\infty$$

$$\sum \left(\gamma^{(t)} \right)^2 < +\infty$$

Alternatively:

$$v^{(t+1)} = v_j^{(t)} + \gamma^{(t+1)} \left(A_t - \text{target} \right)$$