Approximate Bayesian Inference

$$\underbrace{p(\theta \mid x_{1:n})}_{\text{posterior}} \propto \underbrace{p(x_{1:n} \mid \theta)}_{\text{likelihood}} \underbrace{p(\theta)}_{\text{prior}}$$

→ models that have analytical expressions require that

$$\underbrace{p(x_{1:n})}_{\text{marginal likelihood}} = \int p(x_{1:n} \mid \theta)\, p(\theta)\, d\theta$$

$$x_i \sim \text{Bernoulli}(\pi) \qquad p(x_i \mid \pi) = \pi^{x_i}(1-\pi)^{1-x_i}$$

$$\pi \sim \text{Beta}(\alpha_0, \beta_0) \qquad p(\pi) = \frac{\pi^{\alpha_0 - 1}(1-\pi)^{\beta_0 - 1}}{B(\alpha_0, \beta_0)}$$

$$p(x_i) = \int_0^1 p(x_i \mid \pi)\, p(\pi)\, d\pi$$

$$= \int_0^1 \pi^{x_i}(1-\pi)^{1-x_i} \frac{\pi^{\alpha_0 - 1}(1-\pi)^{\beta_0 - 1}}{B(\alpha_0, \beta_0)}\, d\pi$$

$$= \frac{1}{B(\alpha_0, \beta_0)} \int_0^1 \pi^{\overbrace{x_i + \alpha_0 - 1}^{\alpha'}} (1-\pi)^{1 - x_i + \overbrace{\beta_0 - 1}^{\beta'}} d\pi$$

$$= \frac{1}{B(\alpha_0, \beta_0)} \int_0^1 \frac{B(\alpha', \beta')}{B(\alpha', \beta')} \pi^{\alpha'-1} (1-\pi)^{\beta'-1} d\pi$$

$$= \frac{B(\alpha', \beta')}{B(\alpha_0, \beta_0)}$$

But, what if we can't compute $p(x_{1:n})$?

(1) Parametric approximations

Find q closest as possible to $p(\theta | x_{1:n})$

Laplace Approximation ← most popular approach

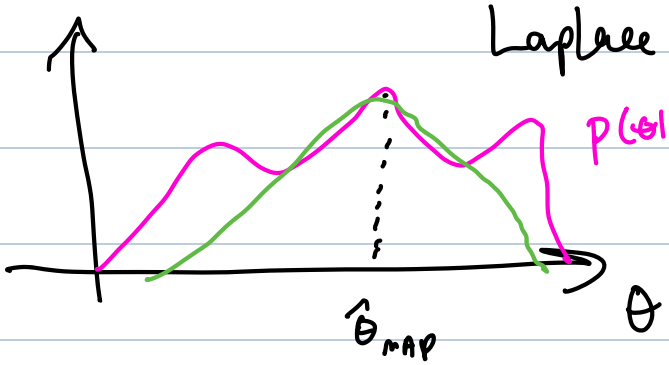
$$\theta | X_{1:n} \dot{\sim} N\left(\hat{\theta}_{MAP}, I_B^{-1}(\hat{\theta}_{MAP})\right)$$

posterior mode: $\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta | X_{1:n})$

$$I_B(\hat{\theta}_{MAP}) = -\mathbb{E}\left[\frac{\partial^2 \log p(\theta | X_{1:n})}{\partial \theta^2} \,\middle|\, \theta = \hat{\theta}_{MAP}\right]$$

Limitations:

- If θ is multivariate, then you need the Hessian matrix of $\log p(\theta | X_{1:n})$
  - → expensive to compute in high dimensions
- Asymptotic normality is for large n. If n is small, the Laplace approximation will be inaccurate
- If $p(\theta | x_{1:n})$ is multimodal, the Laplace approximation will fail



Monte Carlo Methods

→ backed by SLLN:

Given a random sample $X_1, \dots, X_n$ with population $p(x)$, under the assumption $\mathbb{E}[X] < \infty$ then

$X_1, \dots, X_n \longrightarrow \mathbb{E}[X] < \infty$

$Y_1 = g(X_1), \dots, Y_n = g(X_n) \longrightarrow \mathbb{E}[Y] < \infty$

$$\overline{X}_n \xrightarrow{a.s.} \mathbb{E}[X]$$

→ Idea of Monte Carlo is to use samples from $p(\theta \mid x_{1:n})$ as a proxy for the distribution to compute any expectation values we want

- Standard Monte Carlo: Assumes we can sample from $\pi(\theta) \triangleq p(\theta \mid x_{1:n})$

$$\theta^{(1)}, \dots, \theta^{(M)} \overset{i.i.d.}{\sim} \pi$$

$$\mathbb{E}_{\pi}[h(\theta)] \approx \frac{1}{M} \sum_{m=1}^{M} h(\theta^{(m)})$$

$$\xrightarrow{a.s.} \int h(\theta)\, \pi(\theta)\, d\theta$$

→ Standard Monte Carlo assumes we can sample from the posterior, which is generally not the case

Rejection Sampling: $\pi(\theta) \triangleq p(\theta \mid x_{1:n})$

$\tilde{\pi}(\theta) \triangleq p(x_{1:n} \mid \theta)\, p(\theta)$

→ Idea: Sample θ from another distribution $q(\theta)$ and reject some of those samples with some probability

Algorithm: Requires $q(\theta)$, $M = \sup_{\theta} \frac{\tilde{\pi}(\theta)}{q(\theta)}$

$i = 1$

While (keep_sampling):

1. Sample $\theta_* \sim q(\theta)$
2. Sample $u \sim \mathcal{U}(0, 1)$

3. If $u \leq \frac{\tilde{\pi}(\theta)}{M q(\theta)}$:

θ_* is accepted as a sample

$\theta^{(i)} = \theta_*$

$i = i+1$

$\{\theta^{(i)}\}_i$

We can show that set of samples $\{\theta^{(i)}\}$ are samples from $\pi(\theta)$:

Let $\tilde{\theta}$ be a random generated from the algo.

$$F_{\tilde{\theta}}(a) = \mathbb{P}(\tilde{\theta} \leq a)$$

$$= \mathbb{P}\left(\theta_* \leq a \mid u \leq \frac{\tilde{\pi}(\theta_*)}{M q(\theta_*)}\right)$$

$$\overset{(1.)}{=} \frac{\mathbb{P}\left(\theta_* \leq a, \; U \leq \frac{\tilde{\pi}(\theta_*)}{M q(\theta_*)}\right)}{\overset{(2.)}{\mathbb{P}\left(U \leq \frac{\tilde{\pi}(\theta_*)}{M q(\theta_*)}\right)}}$$

1.

$$\mathbb{P}\left(\theta_* \le a,\ U \le \frac{\tilde{\pi}(\theta_*)}{M q(\theta_*)}\right) =$$

Law of total prob.

$$\int \mathbb{P}\left(\theta_* \le a,\ U \le \frac{\tilde{\pi}(\theta_*)}{M q(\theta_*)} \;\middle|\; \theta_* = \theta\right) q(\theta)\, d\theta$$

$$= \int_{-\infty}^{a} \mathbb{P}\left(U \le \frac{\tilde{\pi}(\theta_*)}{M q(\theta_*)} \;\middle|\; \theta_* = \theta\right) q(\theta)\, d\theta$$

$$= \int_{-\infty}^{a} \frac{\tilde{\pi}(\theta)}{M q(\theta)}\, q(\theta)\, d\theta = \frac{1}{M} \int_{-\infty}^{a} \tilde{\pi}(\theta)\, d\theta$$

$$Z = \int_{-\infty}^{\infty} \tilde{\pi}(\theta)\, d\theta$$

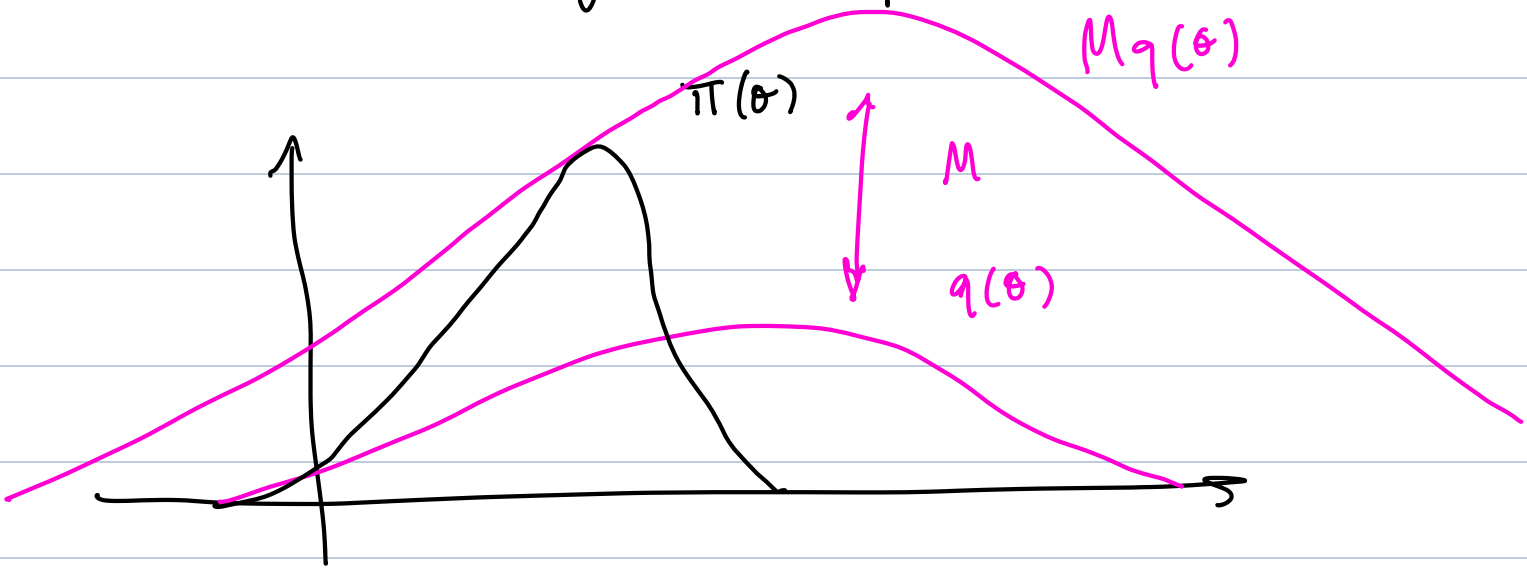
$$= \frac{1}{M} \int_{-\infty}^{a} \frac{Z}{Z} \tilde{\pi}(\theta)\, d\theta = \boxed{\frac{Z}{M}} F_{\theta}(a)$$

Using same approach, show (2.) $= \frac{Z}{M}$

Thus, $F_{\tilde{\theta}}(a) = F_{\theta}(a) \Rightarrow \tilde{\theta} = \theta$ ☒

- Rate of rejection: $\frac{Z}{M} \leftarrow \mathbb{P}\left(U \le \frac{\tilde{\pi}(\theta)}{M q(\theta)}\right)$

- Limitations: If q is poorly chosen, we have to set M to be large and thus we reject too many samples



- Importance Sampling:

If we know $\pi(\theta)$:

$$\mathbb{E}_{\pi}[h(\theta)] = \int h(\theta)\,\pi(\theta)\,d\theta \qquad q(\theta) > 0 \text{ whenever } \pi(\theta) > 0$$

$$= \int h(\theta)\,\frac{\pi(\theta)}{q(\theta)}\,q(\theta)\,d\theta$$

$$= \mathbb{E}_{q}\left[h(\theta)\,w(\theta)\right] \qquad w(\theta) = \frac{\pi(\theta)}{q(\theta)}$$

Importance weight

Normalized Importance Sampling

$$\approx \frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})\,w(\theta^{(m)})$$

$\theta^{(m)} \overset{iid}{\sim} q(\theta)$

$\xrightarrow{\text{a.s.}}$ $\mathbb{E}_\pi[h(\theta)]$ by SLLN

If we know $\tilde{\pi}(\theta)$:

$$\mathbb{E}_\pi[h(\theta)] = \int h(\theta) \frac{\tilde{\pi}(\theta)}{\int \tilde{\pi}(\theta) d\theta} d\theta$$

$\int \tilde{\pi}(\theta) d\theta \triangleq Z$ ← normalizing constant

$$= \frac{1}{Z} \int h(\theta) \frac{\tilde{\pi}(\theta)}{q(\theta)} q(\theta) d\theta$$

$\frac{\tilde{\pi}(\theta)}{q(\theta)} = \tilde{w}(\theta)$

$$= \frac{1}{Z} \mathbb{E}_q[h(\theta)\tilde{w}(\theta)]$$

$$Z = \int \tilde{\pi}(\theta) d\theta = \int \tilde{\pi}(\theta) \frac{q(\theta)}{q(\theta)} d\theta$$

$$= \mathbb{E}_q[\tilde{w}(\theta)] \quad \hat{I}$$

$$\mathbb{E}_\pi[h(\theta)] = \frac{\mathbb{E}_q[h(\theta)\tilde{w}(\theta)]}{\mathbb{E}_q[\tilde{w}(\theta)]} \approx \frac{\frac{1}{M}\sum_{m=1}^{M} h(\theta^{(m)})\tilde{w}(\theta^{(m)})}{\frac{1}{M}\sum_{m=1}^{M} \tilde{w}(\theta^{(m)})}$$

$\xrightarrow{a.s.}$ $\mathbb{E}_{\tilde{\pi}}[h(\theta)]$

→ numerator converges a.s. to $\int \tilde{\pi}(\theta)\, h(\theta)\, d\theta$

→ denominator converges a.s. to $\int \tilde{\pi}(\theta)\, d\theta$

Limitations: Variance of importance sampling estimator scales with the variance of $\tilde{w}(\theta)$

$$\mathbb{V}[\hat{I}] \propto \mathbb{V}[|h(\theta)|\, w(\theta)]$$

→ q should be close to $|h(\theta)|\pi(\theta)$ to reduce this variance

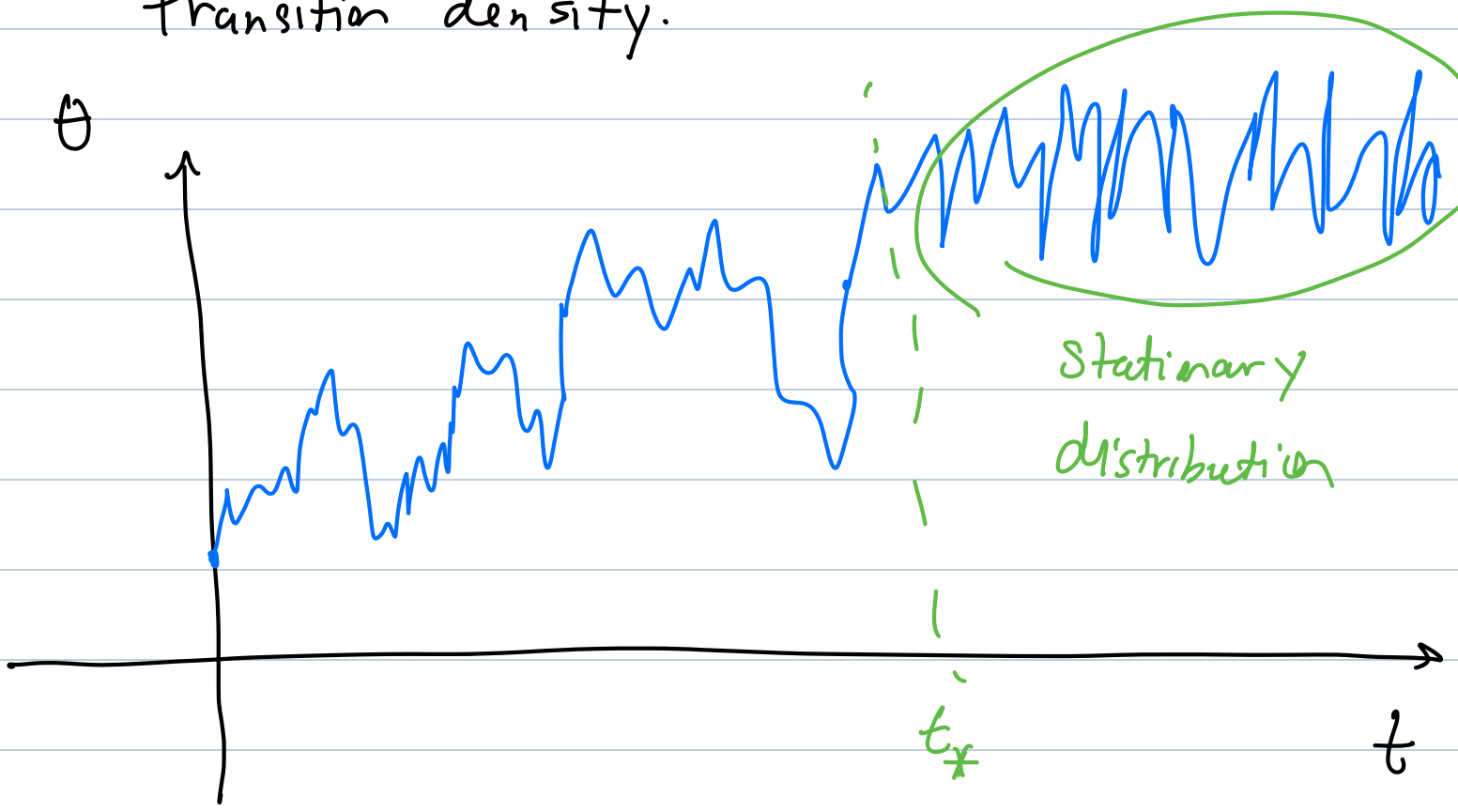
Markov chain Monte Carlo (MCMC)

A stochastic process is an indexed collection of RVs $\{\theta^{(s)}\}_S$ where S is the index set and it is assumed to be uncountable.

Def. Markov chain is a stochastic process $\{\theta^{(t)}\}$ that satisfies the Markov property

$$p(\theta^{(t+1)} \mid \theta^{(0)}, \ldots, \theta^{(t)}) = p(\theta^{(t+1)} \mid \theta^{(t)})$$

The distribution $p(\theta^{(t+1)} \mid \theta^{(t)})$ is called transition density.



Def. A stationary distribution for the Markov chain $\{\theta^{(t)}\}$ with transition density $T(\theta^{(t+1)} \mid \theta^{(t)})$ has the property

$$p(\theta_*) = \int p(\theta)\, T(\theta \mid \theta_*)\, d\theta$$

MCMC idea: Design an algorithm that constructs a Markov chain with stationary distribution that is $\pi(\theta)$

↳ Existence of stationary distribution

↳ Markov chain should be ergodic

↳ aperiodic chain

↳ positive recurrence $\mathbb{E}[T_{\theta_*}] < \infty$

expected time to revisit a state θ_* is finite

Algorithm: Initialize $\theta^{(0)}$

for $t = 1, \ldots, \infty$:

1. Sample $\theta_* \sim q(\theta \mid \theta^{(i-1)})$

ex: $N(\theta^{(i-1)}, \sigma^2)$

2. Sample $u \sim U(0,1)$

$$\alpha = \frac{\tilde{\pi}(\theta_*)}{\tilde{\pi}(\theta^{(i-1)})} \frac{q(\theta^{(i-1)}|\theta_*)}{q(\theta_*|\theta^{(i-1)})}$$

3. Check if $u \leq \min(\alpha, 1)$:

if true:

$$\theta^{(i)} = \theta_*$$

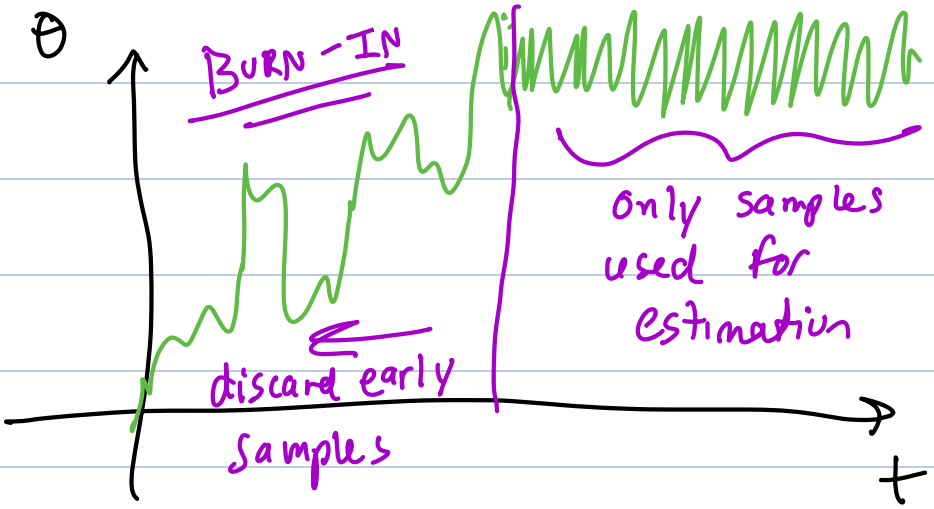
otherwise

$$\theta^{(i)} = \theta^{(i-1)}$$

We can show that this algorithm's constructed Markov chain has stationary distribution = to $\pi(\theta)$

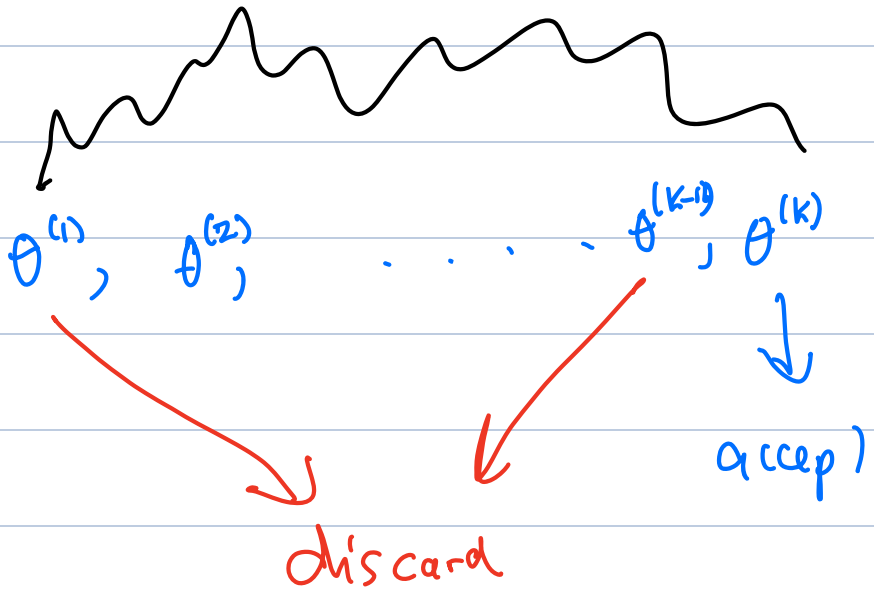
Important aspects:

1. Stationarity is not reached immeadiately, it will be achieved as $t \to \infty$



2. Chains can be auto-correlated in practice:

→ Thinning is applied to Markov chains in practice where only every K^{th} sample is accepted



Linear Models (Chapter 4 - Kay Estimation Theory)

$$\underset{\sim}{x} = \underline{\underline{H}}\,\underset{\sim}{\theta} + \underset{\sim}{w}$$

$\underset{\sim}{x} \in \mathbb{R}^n$, $\underline{\underline{H}} \in \mathbb{R}^{n \times d}$, $\underset{\sim}{\theta} \in \mathbb{R}^d$, $\underset{\sim}{w} \in \mathbb{R}^n$

→ $\underline{\underline{H}}$ is an observation matrix and is assumed to be known

→ $\underset{\sim}{w}$ is zero-mean white noise

$$\mathbb{E}[\underset{\sim}{w}] = 0$$

$$\mathbb{E}[\underset{\sim}{w}\underset{\sim}{w}^T] = \sigma^2 \underline{\underline{I}}$$

identity matrix $n \times n$

Least Squares: Find a point estimator $\hat{\theta}$ that minimizes the squared error between our model's prediction and the observed data

$$\hat{\theta} = \arg\min_{\theta} \mathbb{E}\left[\| \underline{x} - (\underline{\underline{H}}\hat{\theta} + \underline{v}) \|_2^2 \right]$$

take gradient

$$L(\theta) = \| x - H\hat{\theta} \|_2^2 \longrightarrow \frac{\partial L}{\partial \theta} = -2H^T(x - H\hat{\theta}) = 0$$

$$H^T x - H^T H \hat{\theta} = 0 \rightarrow H^T x = H^T H \hat{\theta}$$

$$\boxed{\hat{\theta} = (H^T H)^{-1} H^T x}$$

- What is $\mathbb{E}[\hat{\theta}]$?

$$\mathbb{E}\left[(H^T H)^{-1} H^T x \right] = (H^T H)^{-1} H^T \mathbb{E}[x]$$

$$= \underbrace{(H^T H)^{-1} H^T}\ H\theta$$

$$= \theta$$ Therefore, unbiased estimator

- What is $C_{\hat{\theta}} = \mathbb{E}\left[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T\right]$

$$= \mathbb{E}\left[\hat{\theta}\hat{\theta}^T\right] - \underbrace{\mathbb{E}\left[\hat{\theta}\theta^T\right]}_{\theta\theta^T} - \underbrace{\mathbb{E}\left[\theta\hat{\theta}^T\right]}_{\theta\theta^T} + \underbrace{\mathbb{E}\left[\theta\theta^T\right]}_{\theta\theta^T}$$

$$= \mathbb{E}\left[\hat{\theta}\hat{\theta}^T\right] - \theta\theta^T$$

$$= \mathbb{E}\left[(H^T H)^{-1} H \underbrace{x}_{H\theta + w} x^T H^T (H^T H)^{-1}\right]$$

$$= \mathbb{E}\Big[\underbrace{(H^T H)^{-1} H (H\theta + w)(\theta^T H^T + w^T) H^T (H^T H)^{-1}}\Big]$$

$$\sigma^2 (H^T H)^{-1} + \theta\theta^T$$

$$= \sigma^2 (H^T H)^{-1}$$ ← In the case $p(w)$ is Gaussian, this is the MVUE and achieves the CRLB

Alternative derivation for linear model

$$w \sim N(0, \sigma^2 I)$$
$$x = H\theta + w$$

Then,

$$p(x; \theta) = \mathcal{N}(x; H\theta, \sigma^2 I)$$
$$= (2\pi\sigma^2)^{-n/2} \exp\left(\frac{-1}{2\sigma^2} (x - H\theta)(x - H\theta)^T \right)$$

$$\ell(\theta; x) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|x - H\theta\|_2^2$$

$$\hat{\theta} = \arg\max_{\theta} \ell(\theta; x) = \arg\min_{\theta} \|x - H\theta\|_2^2$$

$$\hat{\theta} = (H^T H)^{-1} H^T x \quad \text{for both cases!}$$

Non-Isotropic Covariance: $w \sim N(0, \underline{\underline{C}})$

$$\underline{\underline{C}}^{-1} = D^T D$$

assume since $\underline{\underline{C}}$ and $\underline{\underline{C}}^{-1}$ by definition are positive definite, they can be factored via Cholesky

$$\mathbb{E}\left[(Dw)(Dw)^T\right] = DCD^T$$
$$= DD^{-1}(D^{-1})^T D^T = I$$

$$X = H\theta + w \quad , \quad \mathbb{E}[ww^T] = C$$

$$DX = DH\theta + Dw$$

$$\tilde{x} = \tilde{H}\theta + \tilde{w}$$

$$\tilde{x} = Dx$$
$$\tilde{H} = DH$$
$$\tilde{w} = Dw$$

$$\hat{\theta} = (\tilde{H}^T\tilde{H})^{-1}\tilde{H}^T\tilde{x}$$

$$= (H^T D^T DH)^{-1} H^T D^T Dx$$