Formula Sheet for the Midterm

ELG 5218 Uncertainty Evaluation in Engineering Measurements and Machine Learning

Instructor: Miodrag Bolić, University of Ottawa

1 Core Bayesian Identities

1.1 Bayes' Theorem

$$p(\theta \mid D) = \frac{p(D \mid \theta)\, p(\theta)}{p(D)}, \tag{1}$$

$$p(D) = \int p(D \mid \theta)\, p(\theta)\, d\theta. \tag{2}$$

Posterior is often written up to proportionality:

$$p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta). \tag{3}$$

1.2 Likelihood for IID Data

For IID data $D = \{x_i\}_{i=1}^N$:

$$p(D \mid \theta) = \prod_{i=1}^{N} p(x_i \mid \theta), \tag{4}$$

$$\log p(D \mid \theta) = \sum_{i=1}^{N} \log p(x_i \mid \theta). \tag{5}$$

1.3 MLE and MAP

Maximum Likelihood Estimate (MLE):

$$\hat{\theta}_{\text{MLE}} = \arg\max_{\theta}\ \log p(D \mid \theta). \tag{6}$$

Maximum A Posteriori (MAP):

$$\hat{\theta}_{\text{MAP}} = \arg\max_{\theta}\ \log p(D \mid \theta) + \log p(\theta). \tag{7}$$

1.4 Posterior Predictive Distribution

$$p(x_* \mid D) = \int p(x_* \mid \theta)\, p(\theta \mid D)\, d\theta. \tag{8}$$

A plug-in (approximate) predictive:

$$p(x_* \mid D) \approx p(x_* \mid \hat{\theta}). \tag{9}$$

1.5 Uncertainty Quantification and Intervals

1.5.1 Credible Interval (Equal-Tailed)

For parameter θ with posterior $p(\theta \mid D)$, a $(1-\alpha)$ central credible interval is

$$\text{CrI}_{1-\alpha} = [q_{\alpha/2}, q_{1-\alpha/2}], \tag{10}$$

where q_p is the p-quantile of $p(\theta \mid D)$.

1.5.2 Highest Posterior Density (HPD) Region

HPD region H with mass $(1-\alpha)$ satisfies

$$\int_H p(\theta \mid D)\, d\theta = 1 - \alpha, \tag{11}$$

$$p(\theta \mid D) \geq c,\ \forall \theta \in H, \tag{12}$$

for some threshold c chosen to achieve mass $1-\alpha$.

2 Beta–Binomial Model

2.1 Model

$$k \mid \theta \sim \text{Binomial}(n, \theta), \tag{13}$$

$$\theta \sim \text{Beta}(\alpha, \beta). \tag{14}$$

2.2 Posterior

$$\theta \mid k, n \sim \text{Beta}(\alpha + k,\ \beta + n - k). \tag{15}$$

2.3 Moments and MAP

Posterior mean:

$$\mathbb{E}[\theta \mid k, n] = \frac{\alpha + k}{\alpha + \beta + n}. \tag{16}$$

Posterior variance:

$$\text{Var}[\theta \mid k, n] = \frac{(\alpha + k)(\beta + n - k)}{(\alpha + \beta + n)^2(\alpha + \beta + n + 1)}. \tag{17}$$

MAP estimator (for $\alpha, \beta > 1$):

$$\hat{\theta}_{\text{MAP}} = \frac{k + \alpha - 1}{n + \alpha + \beta - 2}. \tag{18}$$

3 Gaussian Distributions and Identities

3.1 Scalar Gaussian

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{19}$$

Precision form with $\lambda = 1/\sigma^2$:

$$p(x) \propto \sqrt{\lambda} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right). \tag{20}$$

3.2 Multivariate Gaussian

$$p(z) = \mathcal{N}(z; \mu, \Sigma) \tag{21}$$

$$= \frac{1}{(2\pi)^{D/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(z-\mu)^\top \Sigma^{-1}(z-\mu)\right). \tag{22}$$

Precision form with $\Lambda = \Sigma^{-1}$:

$$p(z) \propto |\Lambda|^{1/2} \exp\left(-\frac{1}{2}(z-\mu)^\top \Lambda (z-\mu)\right). \tag{23}$$

3.3 Partitioned Gaussian: Marginal and Conditional

Joint Gaussian:

$$\begin{bmatrix} x \\ y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix} \right). \tag{24}$$

Marginal of x:

$$x \sim \mathcal{N}(\mu_x, \Sigma_{xx}). \tag{25}$$

3.4 Theorem 2: Conditional of x given y

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}), \tag{26}$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \tag{27}$$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}. \tag{28}$$

This is Theorem 2 (conditioning) expressed with (x, y).

3.5 Theorem 3: Affine Transformation

Assume

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \tag{29}$$

$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_{y|x}), \tag{30}$$

where the conditional $\boldsymbol{y} \mid \boldsymbol{x}$ is a linear function of $\boldsymbol{x}$ plus Gaussian noise. Then the stacked vector $(\boldsymbol{x}, \boldsymbol{y})^\top$ is jointly Gaussian:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b} \end{pmatrix}, \boldsymbol{R} \right), \tag{31}$$

with block covariance

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x\boldsymbol{A}^\top \\ \boldsymbol{A}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{y|x} + \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top \end{pmatrix}. \tag{32}$$

Corollary 2: Marginal of y (Predictive Distribution)

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \tag{33}$$

where

$$\boldsymbol{\mu}_y = \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b}, \tag{34}$$

$$\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_{y|x} + \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top. \tag{35}$$

Corollary 1: Posterior of x given y
$p(\boldsymbol{x} \mid \boldsymbol{y})$:

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}), \tag{36}$$

with

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^\top \boldsymbol{\Sigma}_{y|x}^{-1} \boldsymbol{A} \right)^{-1}, \tag{37}$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y} \left(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \boldsymbol{A}^\top \boldsymbol{\Sigma}_{y|x}^{-1}(\boldsymbol{y} - \boldsymbol{b}) \right). \tag{38}$$

3.6 Product and Ratio of Scalar Gaussians

Product: if

$$\mu_1(x) \propto \mathcal{N}(x; m_1, \sigma_1^2), \tag{39}$$

$$\mu_2(x) \propto \mathcal{N}(x; m_2, \sigma_2^2), \tag{40}$$

then

$$\mu_1(x)\mu_2(x) \propto \mathcal{N}(x; m, \sigma^2), \tag{41}$$

$$\sigma^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \right)^{-1}, \tag{42}$$

$$m = \sigma^2 \left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2} \right). \tag{43}$$

Ratio: with same μ_1, μ_2,

$$\frac{\mu_1(x)}{\mu_2(x)} \propto \mathcal{N}(x; \tilde{m}, \tilde{\sigma}^2), \tag{44}$$

$$\tilde{\sigma}^2 = \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right)^{-1}, \tag{45}$$

$$\tilde{m} = \tilde{\sigma}^2 \left(\frac{m_1}{\sigma_1^2} - \frac{m_2}{\sigma_2^2} \right). \tag{46}$$

4 Gaussian Models and Conjugate Updates

4.1 4.1 Known Variance, Unknown Mean (Scalar)

Model: $x_i \mid \mu, \sigma^2$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$, known σ^2, precision $\lambda = 1/\sigma^2$.

Uniform Prior $p(\mu) \propto 1$:

$$\mu \mid x_{1:n} \sim \mathcal{N}\left(\bar{x},\ \frac{\sigma^2}{n} \right), \tag{47}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

Normal Prior $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ with precision $\lambda_0 = 1/\sigma_0^2$:

$$\lambda_n = \lambda_0 + n\lambda, \tag{48}$$

$$\mu_n = \frac{\lambda_0\mu_0 + n\lambda\bar{x}}{\lambda_n}, \tag{49}$$

$$\mu \mid x_{1:n} \sim \mathcal{N}\left(\mu_n,\ \frac{1}{\lambda_n} \right). \tag{50}$$

4.2 Sequential (Online) Update for Mean

Let μ_{n-1}, λ_{n-1} be posterior parameters after $n-1$ points; add x_n.

$$\lambda_n = \lambda_{n-1} + \lambda, \tag{51}$$

$$w_n = \frac{\lambda}{\lambda_n}, \tag{52}$$

$$\mu_n = \mu_{n-1} + w_n(x_n - \mu_{n-1}). \tag{53}$$

4.3 Unknown Mean and Variance: Normal–Gamma Prior

Use precision $\lambda = 1/\sigma^2$.

$$\mu \mid \lambda \sim \mathcal{N}\big(\mu_0, (\kappa_0\lambda)^{-1}\big), \tag{54}$$

$$\lambda \sim \text{Gamma}(\alpha_0, \beta_0). \tag{55}$$

Posterior hyperparameters after n observations with sample mean $\bar{x}$:

$$\kappa_n = \kappa_0 + n, \tag{56}$$

$$\mu_n = \frac{\kappa_0 \mu_0 + n\bar{x}}{\kappa_n}, \tag{57}$$

$$\alpha_n = \alpha_0 + \frac{n}{2}, \tag{58}$$

$$\beta_n = \beta_0 + \frac{1}{2}\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{x} - \mu_0)^2\right). \tag{59}$$

Posterior:

$$p(\mu, \lambda \mid x_{1:n}) = \mathcal{N}\big(\mu; \mu_n, (\kappa_n \lambda)^{-1}\big)\ \mathrm{Gamma}(\lambda; \alpha_n, \beta_n). \tag{60}$$

4.4 Student–t Distribution from Normal–Gamma

If μ, λ have Normal–Gamma posterior as above, the marginal for μ is Student-t:

$$\mu \mid x_{1:n} \sim t_{2\alpha_n}\left(\mu_n,\ \sqrt{\frac{\beta_n}{\alpha_n \kappa_n}}\right). \tag{61}$$

Posterior predictive (new x_*):

$$x_* \mid x_{1:n} \sim t_{2\alpha_n}\left(\mu_n,\ \sqrt{\frac{\beta_n}{\alpha_n}\left(1 + \frac{1}{\kappa_n}\right)}\right). \tag{62}$$

5 Bayesian Linear Regression

5.1 Linear Regression Model

For real-valued targets $y_n \in \mathbb{R}$ and feature vectors x_n:

$$y_n \mid x_n, w \sim \mathcal{N}(w^\top x_n,\ \sigma^2), \tag{63}$$

$$p(y_n \mid x_n, w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n - w^\top x_n)^2\right). \tag{64}$$

Stack all data:

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \qquad X = \begin{bmatrix} x_1^\top \\ \vdots \\ x_N^\top \end{bmatrix}.$$

5.2 Likelihood

$$p(y \mid X, w) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid w^\top x_n, \sigma^2) \tag{65}$$

$$= (2\pi\sigma^2)^{-N/2} \exp\left(-\frac{1}{2\sigma^2}\|y - Xw\|^2\right). \tag{66}$$

5.3 Gaussian Prior on Weights

$$p(w) = \mathcal{N}(w; 0, \lambda^{-1} I), \tag{67}$$

$$\log p(w) = -\frac{\lambda}{2} w^\top w + C. \tag{68}$$

5.4 Posterior (Up to Proportionality)

$$p(w \mid X, y) \propto \exp\left(-\frac{1}{2\sigma^2}\|y - Xw\|^2 - \frac{\lambda}{2} w^\top w\right). \tag{69}$$

5.5 Closed-Form Gaussian Posterior

Posterior precision and covariance:

$$\Sigma_N^{-1} = \lambda I + \frac{1}{\sigma^2} X^\top X, \tag{70}$$

$$\Sigma_N = \left(\lambda I + \frac{1}{\sigma^2} X^\top X\right)^{-1}. \tag{71}$$

Posterior mean:

$$\mu_N = \Sigma_N \left(\frac{1}{\sigma^2} X^\top y\right). \tag{72}$$

Thus the posterior is:

$$p(w \mid D) = \mathcal{N}(w;\ \mu_N,\ \Sigma_N). \tag{73}$$

MAP estimate:

$$w_{\mathrm{MAP}} = \mu_N. \tag{74}$$

5.6 Posterior Predictive Distribution

For a new point x_*:

$$p(y_* \mid x_*, D) = \int \mathcal{N}(y_* \mid w^\top x_*, \sigma^2)\, \mathcal{N}(w; \mu_N, \Sigma_N)\, dw. \tag{75}$$

Closed form:

$$y_* \mid x_*, D \sim \mathcal{N}\Big(\mu_N^\top x_*,\ \ \sigma^2 + x_*^\top \Sigma_N x_*\Big). \tag{76}$$

Predictive mean:

$$\mu_* = \mu_N^\top x_*.$$

Predictive variance:

$$\mathrm{Var}(y_*) = \sigma^2 + x_*^\top \Sigma_N x_*.$$

6 Logistic Regression and Laplace Approximation

6.1 Logistic Model

For binary labels $y_n \in \{0, 1\}$ and features x_n:

$$\sigma(z) = \frac{1}{1 + e^{-z}}, \tag{77}$$

$$p(y_n = 1 \mid x_n, w) = \sigma(w^\top x_n), \tag{78}$$

$$p(y_n = 0 \mid x_n, w) = 1 - \sigma(w^\top x_n). \tag{79}$$

Likelihood (Bernoulli factorization):

$$p(y \mid X, w) = \prod_{n=1}^{N} \sigma(w^\top x_n)^{y_n} \big(1 - \sigma(w^\top x_n)\big)^{1 - y_n}. \tag{80}$$

Gaussian prior on weights:

$$p(w) = \mathcal{N}(w; 0, \alpha^{-1} I). \tag{81}$$

Posterior (up to proportionality):

$$p(w \mid y, X) \propto p(y \mid X, w)\, p(w). \tag{82}$$

6.2 Log-Posterior, Gradient, Hessian

Let $\mu_n = \sigma(w^\top x_n)$ and define $\mu = (\mu_1, \ldots, \mu_N)^\top$. Log-posterior (up to constant):

$$\log p(w \mid D) = -\frac{\alpha}{2} w^\top w + \sum_{n=1}^{N} \Big[y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n) \Big] + C. \tag{83}$$

Gradient:

$$\nabla_w \log p(w \mid D) = -\alpha w + \sum_{n=1}^{N} (y_n - \mu_n) x_n \tag{84}$$

$$= -\alpha w + X^\top (y - \mu). \tag{85}$$

Hessian:

$$H = \nabla_w^2 \log p(w \mid D) \tag{86}$$

$$= -\alpha I - \sum_{n=1}^{N} \mu_n (1 - \mu_n) x_n x_n^\top \tag{87}$$

$$= -(\alpha I + X^\top S X), \tag{88}$$

where S is diagonal with $S_{nn} = \mu_n(1 - \mu_n)$.

6.3 Laplace Approximation for Posterior

Mode (MAP):

$$w_{\text{MAP}} = \arg\max_w \ \log p(w \mid D). \tag{89}$$

Quadratic (Laplace) approximation around w_{MAP}:

$$p(w \mid D) \approx \mathcal{N}(w; w_{\text{MAP}},\ \Sigma_N)\,, \tag{90}$$

$$\Sigma_N^{-1} = -H\big|_{w = w_{\text{MAP}}} = \alpha I + X^\top S X. \tag{91}$$

6.4 Predictive Probability (Approximate)

Exact predictive for new x_*:

$$p(y_* = 1 \mid x_*, D) = \int \sigma(w^\top x_*)\, p(w \mid D)\, dw. \tag{92}$$

Under Laplace approximation $p(w \mid D) \approx \mathcal{N}(w_{\text{MAP}}, \Sigma_N)$ this is a 1D Gaussian integral in $w^\top x_*$ (often evaluated numerically or via standard approximations).

6.5 Potential Energy for Logistic Regression

Define unnormalized posterior $\pi(w) \propto p(y \mid X, w)p(w)$. Potential energy:

$$U(w) = -\log \pi(w) \tag{93}$$

$$= -\sum_{n=1}^{N} \left[y_n \log \mu_n + (1 - y_n) \log(1 - \mu_n) \right] + \frac{\alpha}{2}\, w^\top w + C. \tag{94}$$

Gradient of $U(w)$:

$$\nabla U(w) = X^\top (\mu - y) + \alpha w. \tag{95}$$

7 Markov Chain Monte Carlo (MCMC)

7.1 Metropolis–Hastings Algorithm

Given target density $\pi(\theta)$ and proposal $q(\theta' \mid \theta)$:

$$r(\theta \to \theta') = \frac{\tilde{\pi}(\theta') q(\theta \mid \theta')}{\tilde{\pi}(\theta) q(\theta' \mid \theta)}, \tag{96}$$

$$\alpha(\theta \to \theta') = \min(1, r(\theta \to \theta')), \tag{97}$$

where $\tilde{\pi}$ is any unnormalized version of π. If $\pi(\theta) \propto p(x \mid \theta)p(\theta)$, then

$$r(\theta \to \theta') = \frac{p(x \mid \theta') p(\theta') q(\theta \mid \theta')}{p(x \mid \theta) p(\theta) q(\theta' \mid \theta)}, \tag{98}$$

so the evidence cancels.

7.2 Random-Walk Metropolis (RWM)

Symmetric proposal:

$$q(\theta' \mid \theta) = \mathcal{N}(\theta'; \theta, \sigma_{\text{prop}}^2 I), \tag{99}$$

so acceptance probability simplifies to

$$\alpha(\theta \to \theta') = \min\left(1, \frac{\pi(\theta')}{\pi(\theta)}\right). \tag{100}$$

7.3 Detailed Balance Condition

A sufficient condition for stationarity of π is

$$\pi(\theta) P(\theta' \mid \theta) = \pi(\theta') P(\theta \mid \theta') \tag{101}$$

for all θ, θ'. Metropolis–Hastings transition kernel satisfies detailed balance.

7.4 Monte Carlo Estimation

Given samples $\{\theta^{(s)}\}_{s=1}^S$ from $\pi(\theta)$, an expectation

$$\mathbb{E}_\pi[f(\theta)] = \int f(\theta)\, \pi(\theta)\, d\theta \tag{102}$$

can be approximated as

$$\mathbb{E}_\pi[f(\theta)] \approx \frac{1}{S} \sum_{s=1}^{S} f(\theta^{(s)}). \tag{103}$$

Standard error decays as $\sim 1/\sqrt{S}$ (modulo autocorrelation).

7.5 MCMC Diagnostics

7.5.1 Autocorrelation Function (ACF)

For an MCMC chain $\{\theta_t\}_{t=1}^N$, the autocorrelation at lag ℓ is:

$$\rho(\ell) = \text{Corr}(\theta_t,\ \theta_{t+\ell}). \tag{104}$$

7.5.2 Autocorrelation Time and Effective Sample Size (ESS)

Define the **autocorrelation time**:

$$\tau = 1 + 2\sum_{\ell=1}^{\infty} \rho(\ell). \tag{105}$$

Effective sample size:

$$N_{\text{eff}} = \frac{N}{\tau}. \tag{106}$$

Monte Carlo standard error:

$$\text{SE}(\hat{\theta}) = \frac{\sigma(\theta)}{\sqrt{N_{\text{eff}}}}. \tag{107}$$

7.5.3 Gelman–Rubin Convergence Diagnostic ($\hat{R}$)

Suppose we run J parallel chains, each producing L post–burn-in samples.

Between-chain variance:

$$B = \frac{L}{J-1}\sum_{j=1}^{J} \left(\bar{\theta}_j - \bar{\theta}_\cdot\right)^2, \tag{108}$$

where $\bar{\theta}_j$ is the mean of chain j and $\bar{\theta}_\cdot$ is the mean across chains.

Within-chain variance:

$$W = \frac{1}{J}\sum_{j=1}^{J} s_j^2, \tag{109}$$

where s_j^2 is the sample variance of chain j.

Potential Scale Reduction Factor (PSRF):

$$\hat{R} = \sqrt{\frac{(L-1)W + B}{LW}} = \sqrt{\frac{\text{between-chain variance}}{\text{within-chain variance}}}. \tag{110}$$

8 Langevin Dynamics and MALA

8.1 Langevin SDE

For a target density $p(x)$ with score $\nabla_x \log p(x)$, the Langevin stochastic differential equation is

$$dX_t = \frac{1}{2}\nabla_x \log p(X_t)\, dt + dW_t, \tag{111}$$

where W_t is standard Brownian motion.

8.2 Potential Form

For a target $\pi(\theta) \propto e^{-U(\theta)}$:

$$d\Theta_t = -\frac{1}{2}\nabla U(\Theta_t)\, dt + dW_t. \tag{112}$$

8.3 Unadjusted Langevin Algorithm (ULA)

Euler–Maruyama discretization with step size Δt:

$$\theta_{k+1} = \theta_k - \frac{1}{2}\nabla U(\theta_k)\,\Delta t + \sqrt{\Delta t}\, Z_k, \tag{113}$$

$$Z_k \sim \mathcal{N}(0, I). \tag{114}$$

8.4 Metropolis-Adjusted Langevin Algorithm (MALA)

Langevin proposal:

$$q(\tilde{\theta} \mid \theta) = \mathcal{N}\!\left(\tilde{\theta};\ \theta - \frac{1}{2}\nabla U(\theta)\,\Delta t,\ \Delta t\, I\right). \tag{115}$$

Metropolis–Hastings acceptance probability:

$$\alpha(\theta \to \tilde{\theta}) = \min\left(1, \frac{\pi(\tilde{\theta})\, q(\theta \mid \tilde{\theta})}{\pi(\theta)\, q(\tilde{\theta} \mid \theta)}\right) \tag{116}$$

$$= \min\left(1, \exp\big(-U(\tilde{\theta}) + U(\theta)\big)\frac{q(\theta \mid \tilde{\theta})}{q(\tilde{\theta} \mid \theta)}\right). \tag{117}$$

9 Bayesian Filtering (General State-Space Model)

Density form: $p(z_k \mid z_{k-1})$, $p(y_k \mid z_k)$.

Prediction (Chapman–Kolmogorov):

$$p(z_k \mid y_{1:k-1}) = \int p(z_k \mid z_{k-1})\, p(z_{k-1} \mid y_{1:k-1})\, dz_{k-1}. \tag{118}$$

Update (Bayes' rule):

$$p(z_k \mid y_{1:k}) = \frac{p(y_k \mid z_k)\, p(z_k \mid y_{1:k-1})}{p(y_k \mid y_{1:k-1})}, \tag{119}$$

$$p(y_k \mid y_{1:k-1}) = \int p(y_k \mid z_k) p(z_k \mid y_{1:k-1}) dz_k. \tag{120}$$

Joint factorization (SSM):

$$p(z_{0:T}, y_{1:T}) = p(z_0)\prod_{k=1}^{T} p(z_k \mid z_{k-1})\prod_{k=1}^{T} p(y_k \mid z_k). \tag{121}$$

9.1 Linear–Gaussian State Space Model (SSM)

Dynamics: $z_t = F_t z_{t-1} + B_t u_t + b_t + q_t$, $q_t \sim \mathcal{N}(0, Q_t)$.

$$p(z_t \mid z_{t-1}, u_t) = \mathcal{N}(z_t;\ F_t z_{t-1} + B_t u_t + b_t,\ Q_t). \tag{122}$$

Measurement: $y_t = H_t z_t + D_t u_t + d_t + r_t$, $r_t \sim \mathcal{N}(0, R_t)$.

$$p(y_t \mid z_t, u_t) = \mathcal{N}(y_t;\ H_t z_t + D_t u_t + d_t,\ R_t). \tag{123}$$

10 Kalman Filter (Complete Algorithm)

Initialization: $p(z_0) = \mathcal{N}(z_0; \mu_0, \Sigma_0)$. **For** t=1,...,T:

Prediction:

$$\mu_{t|t-1} = F_t \mu_{t-1|t-1} + B_t u_t + b_t,$$
$$\Sigma_{t|t-1} = F_t \Sigma_{t-1|t-1} F_t^\top + Q_t. \tag{124}$$

Update:

$$\hat{y}_t = H_t \mu_{t|t-1} + D_t u_t + d_t, \qquad S_t = H_t \Sigma_{t|t-1} H_t^\top + R_t, \tag{125}$$

$$K_t = \Sigma_{t|t-1} H_t^\top S_t^{-1}, \qquad \mu_{t|t} = \mu_{t|t-1} + K_t(y_t - \hat{y}_t), \tag{126}$$

$$\Sigma_{t|t} = (I - K_t H_t)\Sigma_{t|t-1}. \tag{127}$$

Innovation (residual):
$\tilde{y}_t = y_t - \hat{y}_t, \qquad \tilde{y}_t \sim \mathcal{N}(0, S_t)$ (consistency check).

10.1 Kalman Filter (1D Scalar Special Case)

Model:

$$z_t = z_{t-1} + q_t,\ q_t \sim \mathcal{N}(0, q_t), \qquad y_t = z_t + r_t,\ r_t \sim \mathcal{N}(0, r_t). \tag{128}$$

Prediction:

$$\mu_{t|t-1} = \mu_{t-1|t-1}, \qquad \sigma^2_{t|t-1} = \sigma^2_{t-1|t-1} + q_t. \tag{129}$$

Update:

$$\hat{y}_t = \mu_{t|t-1}, \qquad S_t = \sigma^2_{t|t-1} + r_t, \qquad K_t = \frac{\sigma^2_{t|t-1}}{\sigma^2_{t|t-1} + r_t}, \tag{130}$$

$$\mu_{t|t} = \mu_{t|t-1} + K_t(y_t - \hat{y}_t), \qquad \sigma^2_{t|t} = (1 - K_t)\sigma^2_{t|t-1}. \tag{131}$$

10.2 Monte Carlo and Importance Sampling

Monte Carlo expectation:

$$\mathbb{E}[f(Z)] = \int f(z)p(z)\,dz \approx \frac{1}{M}\sum_{m=1}^{M} f(z^{(m)}), \tag{132}$$

$$z^{(m)} \sim p(z). \tag{133}$$

Importance sampling identity:

$$\mathbb{E}[f(Z)] = \int f(z)\frac{p(z)}{q(z)}\,q(z)\,dz \approx \sum_{m=1}^{M} \tilde{w}^{(m)} f(z^{(m)}),$$

$$z^{(m)} \sim q(z), \tag{134}$$

$$w^{(m)} = \frac{p(z^{(m)})}{q(z^{(m)})}, \qquad \tilde{w}^{(m)} = \frac{w^{(m)}}{\sum_{j=1}^{M} w^{(j)}}. \tag{135}$$

10.3 Sequential Importance Sampling / Particle Approximation

Empirical approximation:

$$p(z_k \mid y_{1:k}) \approx \sum_{m=1}^{M} w_k^{(m)}\,\delta(z_k - z_k^{(m)}). \tag{136}$$

General SMC weight update (observation-dependent proposal):

$$w_k^{(m)} \propto w_{k-1}^{(m)} \frac{p(y_k \mid z_k^{(m)})\,p(z_k^{(m)} \mid z_{k-1}^{(m)})}{q(z_k^{(m)} \mid z_{k-1}^{(m)}, y_k)}. \tag{137}$$

Bootstrap / prior proposal ($q = p(z_k \mid z_{k-1})$):

$$w_k^{(m)} \propto w_{k-1}^{(m)}\,p(y_k \mid z_k^{(m)}). \tag{138}$$

10.4 Resampling (SIR / PF)

Resampling distribution: sample indices from

$$\sum_{j=1}^{M} w_k^{(j)}\,\delta(z_k - z_k^{(j)}), \tag{139}$$

then set new weights to $1/M$:

$$\{\bar{z}_k^{(m)}\}_{m=1}^{M} \sim \sum_{j=1}^{M} w_k^{(j)}\delta(z_k - z_k^{(j)}), \qquad \bar{w}_k^{(m)} = \frac{1}{M}. \tag{140}$$

State estimate (posterior mean):

$$\hat{z}_k = \sum_{m=1}^{M} w_k^{(m)} z_k^{(m)}. \tag{141}$$

10.5 Optimal Proposal (Variance-Minimizing)

Optimal proposal:

$$q_k^{\text{opt}}(z_k \mid z_{k-1}, y_k) = p(z_k \mid z_{k-1}, y_k). \tag{142}$$

Weight simplification under q^{opt}:

$$w_k^{(m)} \propto w_{k-1}^{(m)}\,p(y_k \mid z_{k-1}^{(m)}), \tag{143}$$

$$p(y_k \mid z_{k-1}) = \int p(y_k \mid z_k)p(z_k \mid z_{k-1})\,dz_k. \tag{144}$$