MALA and Barker's Proposal: Gradient-Based MCMC Methods

From RWM to more advanced methods

Random Walk Metropolis (RWM):

$$q^* = q + \sigma W, \quad W \sim N(0, I_d)$$

Fundamental Trade-off:

- Large step-size σ: Low acceptance
- Small σ: Slow exploration
- Optimal: $\sigma = \mathcal{O}(d^{-1})$

Problem: In high dimensions, RWM becomes inefficient

- Optimal acceptance rate: 0.234
- Curse of dimensionality: step size $\propto 1/d$

From Langevin Diffusion to MALA

Continuous Langevin Diffusion:

$$dX_t = \frac{1}{2}\nabla \log \pi(X_t)dt + dB_t$$

- Has π as stationary distribution
- Gradient provides moves toward high-probability regions

Euler-Maruyama Discretization (ULA):

$$X^{(t)} = X^{(t-1)} + \frac{\epsilon}{2}\nabla \log \pi(X^{(t-1)}) + \sqrt{\epsilon}W$$

Problem π is **not** the invariant distribution of ULA!
Solution: Add Metropolis-Hastings correction $\Rightarrow$ MALA

Metropolis-Adjusted Langevin Algorithm

just for reference. do not write down algorithm during exam

Algorithm 1 MALA

Input: Initial $X^{(0)}$

for $t = 1, 2, \ldots$ **do**

Propose: $X^* = X^{(t-1)} + \frac{\epsilon}{2}\nabla \log \pi(X^{(t-1)}) + \sqrt{\epsilon}W$

Compute acceptance ratio:

$$\alpha = \min\left\{1, \frac{\pi(X^*)q(X^{(t-1)}|X^*)}{\pi(X^{(t-1)})q(X^*|X^{(t-1)})}\right\}$$

Accept $X^{(t)} = X^*$ with probability α, else $X^{(t)} = X^{(t-1)}$

end for

Optimal Scaling Theory

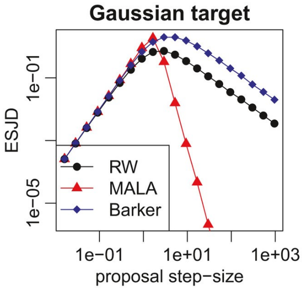
Maximizing Expected Squared Jump Distance (ESJD) $\mathbb{E}\left[\|X^{(t+1)} - X^{(t)}\|^2\right]$

Dimension Scaling:

- RWM: $\sigma = \mathcal{O}(d^{-1})$
- MALA: $\sigma = \mathcal{O}(d^{-1/3})$

Optimal Acceptance:

- RWM: 0.234
- MALA: 0.574



Implication: MALA maintains larger step sizes in high dimensions

- Better exploration efficiency
- Faster convergence to target distribution
- **Catch** - requires gradient computation

From Local-Balanced Proposals to Local-Informed and back again

General idea: Start with symmetric proposal $K_\sigma(x, y) = K_\sigma(y, x)$ with step-size σ and create informed proposal

$$Q_\pi(x, y) \propto \pi(y) K_\sigma(x, y)$$

that bias the proposal toward high-probability states. For large σ $K_\sigma(x, y) \approx$ Uniform and when σ is tiny, $K_\sigma(x, z) \approx 0$ except when z is near x and in that case $\pi(x) \approx \pi(z)$. General form for informed proposals:

$$Q_g(x, y) = \frac{g\left(\frac{\pi(y)}{\pi(x)}\right) K(x, y)}{Z_g(x)}$$

Remarkable result:, Q_g is **locally balanced** if $g(t) = tg(1/t)$.

Barker's Proposal

Expand $\pi(y)/\pi(x)$ around x with local (first-order) approximation:

$$e^{\log \pi(y)} \approx e^{\log \pi(x) + (\nabla \log \pi(x)^T)(y-x)}$$

and use $g(t) = t/(1+t)$. This gives $Z_g(x) = 1/2$ and **Proposal Density:**

$$Q_B(x, dy) = \frac{2}{1 + e^{-(\nabla \log \pi(x))^T (y-x)}} K(x, dy)$$

where $K(x, dy)$ is a base kernel (e.g., Gaussian)
Key Idea: Use gradient to stochastically bias proposal direction

---

Algorithm 2 1D case with Gaussian kernel

Sample $Z \sim N(0, \sigma^2)$
Calculate $p(x, z) = 1/(1 + \exp(-Z^T \nabla \log \pi(x)))$:
Set $b(x, z) = 1$ with probability $p(x, z)$, else $b(x, z) = -1$
Propose $Y = x + b(x, z)Z$

Just for my own reference

This help to understand why Barker proposal use gradient to stochastically bias proposal direction.

Weight of proposal:

$$w = \frac{1}{1 + e^{-(\nabla \log \pi(x))^T (y-x)}} = \frac{1}{1 + e^{-g(y-x)}}$$

where $g = \nabla \log \pi(x)$.

Behavior analysis: Consider four scenarios based on the signs of g and $(y - x)$:

Scenario	$(y-x)$	$g(y-x)$	$e^{-g(y-x)}$	**Weight**	**Meaning**
$g > 0$, move right	> 0	> 0	≈ 0	≈ 1	Favored
$g > 0$, move left	< 0	< 0	$\to \infty$	≈ 0	Penalized
$g < 0$, move left	< 0	> 0	≈ 0	≈ 1	Favored
$g < 0$, move right	> 0	< 0	$\to \infty$	≈ 0	Penalized

Table: Barker proposal weight behavior

MALA vs Barker's Proposal

Aspect	MALA	Barker
Proposal	$Y = x + \frac{\sigma^2}{2}\nabla \log \pi(x) + \sigma Z$	$Y = x \pm Z$ with directional prob
Gradient use	Drift term (deterministic shift)	Direction selection (probabilistic)
Robustness	Sensitive to step size	More robust to large gradients
Scaling	$\mathcal{O}(d^{-1/3})$	$\mathcal{O}(d^{-1/3})$