$\pi \propto \exp(-V)$

MCMC: $X_1, \dots, X_T \sim \hat{\pi} \approx \pi$

$$\hat{\mu}_\pi = \frac{1}{T} \sum X_t \quad , \quad \hat{\Sigma}_\pi = \frac{1}{T} \sum (X_t - \hat{\mu}_\pi) \otimes (X_t - \hat{\mu}_\pi)$$

MCMC as discrete Langevin

$$dX_t = -\nabla V(X_t)\, dt + \sqrt{2}\, dB_t$$

has π as unique stationary dist.

V is strongly convex $\Rightarrow$ approximately sample π after $O(d)$ queries to ∇V

Variational inference:

$$\hat{\pi} = \arg\min_{p \in \mathcal{P}} KL(p \,\|\, \pi)$$

What is $\mathcal{P}$

(1) $\mathcal{P}_2(\mathbb{R}^d)$ - measures with bounded second moments

(2) $\mathcal{P}$ is a product measure - mean-field

(3) p is a Gaussian

How does $\hat{\pi}$ approximate π?

$(m,\Sigma) \mapsto KL(\phi_{(m,\Sigma)} \| \pi)$ not convex

$(m,\Sigma) \mapsto KL(\phi_{(m,\Sigma)} \| \pi)$ not convex in Fisher geometry

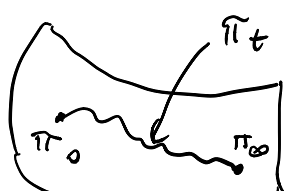
---

Wasserstein gradient flows -

$(\pi_t)_{t \geq 0}$ is the marginal dist. of Langevin diffusion.

$\pi_t = \mathrm{Law}(X_t)$, $dX_t = -\nabla V(X_t)\,dt + \sqrt{2}\, dB_t$

$\partial_t \pi_t = -\mathrm{div}(\pi_t \nabla V) + \Delta \pi_t$ Fokker-Planck



$\propto e^{-V}$

$$W_2(\mu,\nu) = \left[\inf_{\gamma \in C(\mu,\nu)} \int \|x-y\|^2 \, d\gamma(x,y)\right]^{1/2}$$

$(\mathcal{P}_2(\mathbb{R}^d), W_2)$ is a metric space.

Thm (JKO '98):

The law $(\pi_t)_{t\geq 0}$ of the Langevin diffusion is a gradient flow of $KL(\cdot \| \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$.

$$\partial_t(\pi_t) = \operatorname{div}(\pi_t \nabla V) + \Delta \pi_t \iff \dot{X}_t = -\nabla_{W_2} KL(\pi_t \| \pi)(X_t)$$

$$\dot{X}_t = -\nabla V(X_t) - \nabla \log \pi_t(X_t)$$

JKO-scheme

$$\pi_{h,k+1} := \arg\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \left[KL(\mu \| \pi) + \frac{1}{2h} W_2^2(\mu, \mu_{h,k}) \right]$$

now: π_t is unknown so how do you implement this?

Idea: evolve N particles $X_t^{(1)}, \dots, X_t^{(N)}$ and somehow take an expection.

.. -

$$K(x,y) = \begin{bmatrix} e^{-\|x_1 - y_1\|^2/\sigma^2} \\ \vdots \\ \vdots \\ e^{-\|x_d - y_d\|^2/\sigma^2} \end{bmatrix}$$

$$\dot{X}_t = \nabla_{W_2} KL(\pi_t \| \pi)(X_t)$$
$$= -\nabla V(X_t) - \nabla \log \pi_t(X_t)$$

$$\nabla \log \pi_t(x) \approx \int K(x,y) \nabla \log \pi_t(y) \, d\pi_t(y)$$

$$= \int K(x,y) \frac{\nabla \pi_t(y)}{\pi_t(y)} \pi_t(y) dy$$

$$= \int K(x,y) \nabla \pi_t(y) \, dy$$

$$= -\int \nabla_y K(x,y) \, \pi_t(y) dy$$
$$= -\mathbb{E}_{X_t \sim \pi_t} [\nabla_y K(X, X_t)]$$

Project Wasserstein gradient onto RKHS.

$$\nabla \log \pi_t(x) \approx -\frac{1}{N} \sum_{i=1}^{N} \nabla_y K(x, x_t^{(i)})$$

$$\dot{X}_t^{(j)} = -\nabla V(X_t^{(j)}) + \frac{1}{N}\sum_{i=1}^{N} \nabla_y K(X_t^{(i)}, X_t^{(j)})$$

$$, j = 1, \dots, N$$

Some Issues:

1) K?
2) Does not scale for $d > 3$
3) VI with unclear $\mathcal{D}$

Will come back to this

$(\pi_t)_{t \geq 0}$

$\pi_t = \mathrm{Law}(X_t)$ $\qquad$ $dX_t = -\nabla V(X_t)\,dt + \sqrt{2}\,dB_t$

$m_t = \mathbb{E}(X_t)$, $\quad \Sigma_t = \mathrm{cov}(X_t)$

$$\dot{m}_t = -\mathbb{E} V(X_t)$$

$$\dot{\Sigma}_t = 2 I_d - \mathbb{E}\big(\nabla V(X_t) \otimes (X_t - m_t) + (X_t - m_t) \otimes V(X_t)\big)$$

$N(m_t, \Sigma_t)$?

how to compute expectations ?

$X_t \sim \pi_t$ $\qquad$ $Y_t \sim N(m_t, \Sigma_t)$

$$\dot{m}_t = -\mathbb{E}\, \nabla V(Y_t)$$

$$\dot{\Sigma}_t = 2 I_d - \mathbb{E}\big(\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes V(Y_t)\big)$$

$p_t \sim N(m_t, \Sigma_t)$ $\qquad$ $(P2)$

$$m_t \neq \mathbb{E}(X_t), \quad \Sigma_t \neq \mathrm{Cov}(X_t)$$

Thm (L-C-B-B-R '22):
The law $(p_t)_{t\geq 0}$ of (P2) is a gradient flow of $KL(\cdot \| \pi)$ on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ which is constrained to the space of Gaussians.

Bures-Wasserstein space:

$BW(\mathbb{R}^d)$ - $m \in \mathbb{R}^d$

$\Sigma \in S_{++}$ - cone of p.d. matrices

Transport: $p_0 := p_{m_0, \Sigma_0} \to p_1 : p_{m_1, \Sigma_1}$

$$\nabla \varphi(x) = m_1 + \Sigma_0^{-1/2} (\Sigma_0^{1/2} \Sigma_t \Sigma_0^{1/2}) \Sigma_0^{-1/2} (x - m_0)$$

= affine map

$$W_2^2(p_0, p_1) = \|m_1 - m_2\|^2 + tr(\Sigma_0 + \Sigma_1 - 2(\Sigma_0^{1/2} \Sigma_1 \Sigma_0^{1/2})^{\frac{1}{2}})$$

Bures-JKO scheme:

$$(B) \quad p_{n,k+1} := \arg\min_{p \in BW(\mathbb{R}^d)} \left[KL(p \| \pi) + \frac{1}{2} W_2^2(p, p_{n,k}) \right]$$

Thm (L-L-B-B-R '22):
Let $\pi \propto \exp(-V)$ be the target density on $\mathbb{R}^d$. The limiting curve $(\pi_t)_{t \geq 0}$ where $p_t = N(m_t, \Sigma_t)$ is obtained by (B). BW gradient flow $(p_t)_{t \geq 0}$ of the KL divergence $KL(\cdot \| \pi)$ satisfies (P2).

Convergence properties.

$\mathcal{F} : BW(\mathbb{R}^d) \to \mathbb{R} \cup \{\infty\}$ and $\alpha \in \mathbb{R}$

$\mathcal{F}$ is α-convex if for all constant speed geodesics $(p_t)_{t \geq 0}$ in $BW(\mathbb{R}^d)$

$$\mathcal{F}(p_t) \leq (1-t)\,\mathcal{F}(p_0) + t\,\mathcal{F}(p_1) - \frac{\alpha t(1-t)}{2} W_2^2(p_0, p_1)$$

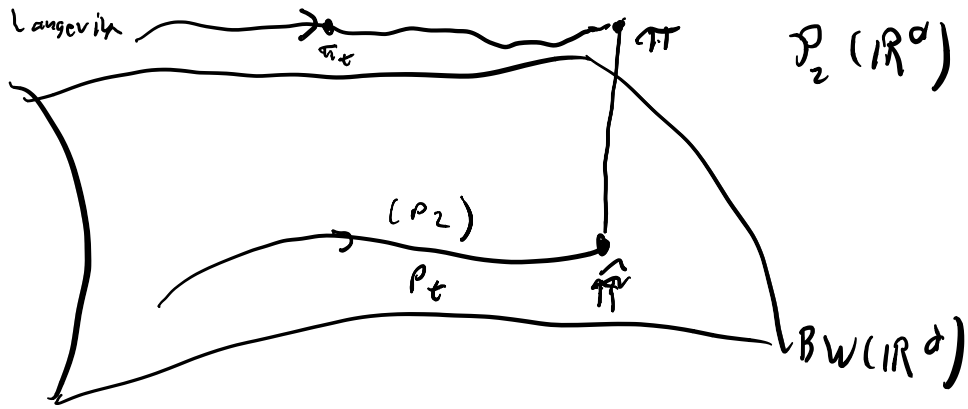
$t \in [0,1]$.

(Lem) For any $\alpha \in \mathbb{R}$ if $\nabla^2 V \succeq \alpha I$ then $KL(\cdot \| \pi)$ is α-convex on $BW(\mathbb{R}^d)$ and there is a unique soln. to BW-gradient flow of $KL(\cdot \| \pi)$ starting at ρ_0 and

1. If $\alpha > 0$, $\forall t \geq 0$
$$W_2^2(\rho_t, \hat{\pi}) \leq \exp(-2\alpha t) W_2^2(\rho_0, \hat{\pi})$$

2. If $\alpha > 0$, $\forall t \geq 0$
$$KL(\rho_t \| \pi) - KL(\hat{\rho} \| \pi) \leq \exp(-2\alpha t) \left[KL(\rho_0 \| \pi) - KL(\hat{\pi} \| \pi) \right]$$

3. If $\alpha = 0$, $\forall t > 0$
$$KL(\rho_t \| \pi) - KL(\hat{\pi} \| \pi) \leq \frac{1}{2t} W_2^2(\rho_0, \hat{\pi}).$$



JKO: store (π_t)

B-JKO: store (m_t, Σ_t)

Algorithms -

1) $$\dot{m}_t = -\mathbb{E}\nabla V(Y_t)$$

$$\dot{\Sigma}_t = 2I_d - \mathbb{E}\big[\nabla V(Y_t) \otimes (Y_t - m_t) + (Y_t - m_t) \otimes \nabla V(Y_t)\big]$$

Practically good

2) Provable algorithm - PA

$\alpha > 0$, $h > 0$, m_0, Σ_0

For $k = 1, \dots, N$

$$\hat{X}_k \sim p_k$$

$$m_{k+1} \leftarrow m_k - h \nabla V(\hat{X}_k)$$

$$M_k \leftarrow I - h\left(\nabla^2 V(\hat{X}_k) - \Sigma_k^{-1}\right)$$

$$\Sigma_k^+ \leftarrow M_k \Sigma_k M_k$$

$$\Sigma_k = \leftarrow \text{clip}^{1/\alpha}(\Sigma_k^+)$$

$$\text{clip}^{\tau}(\Sigma) = \sum_{i=1}^{d} (\lambda_i \wedge \tau) u_i u_i^T$$

Thm: Assume $0 < \alpha I < \nabla^2 V < I$, $h < \frac{\alpha}{6}$, initialize with $\frac{\alpha}{4} I < \Sigma_{\mu_0} < \frac{1}{\alpha} I$, then $\forall$ $k \in \mathbb{N}$

$$\mathbb{E}\, W_2^2(p_k, \hat{\pi}) \le \exp(-\alpha k h)\, W_2^2(p_0, \hat{\pi}) + \frac{21 d h}{\alpha^2}.$$

$$\mathbb{E}\, W_2^2(\rho_K, \hat{\pi}) \le \varepsilon^2 \quad \text{provided}$$

$$h \asymp \frac{\alpha^2 \varepsilon^2}{d} \quad \& \quad K \gtrsim \frac{d}{\alpha^2 \varepsilon^2} \log\left(W_2(\rho_0, \hat{\pi})/\varepsilon\right)$$

Complexity: $\mathcal{O}\left(\frac{d}{\alpha^2 \varepsilon^2} \log\left(\frac{W_2(\rho_0, \hat{\pi})}{\varepsilon}\right)\right)$

PA is SGD

Recall: $(p_t = p_{m_t, \Sigma_t})_{t \ge 0}$

Bures-Wasserstein gradient:

$$g_p := \nabla_{BW} \mathrm{KL}(\cdot \,\|\, \pi)$$

$$= \left(\mathbb{E}_p \nabla V,\ \mathbb{E}_p \nabla^2 V - \mathrm{cov}_p^{-1}\right)$$

$$\hat{g}_p := \left(\nabla V(\hat{X}),\ \nabla^2 V(\hat{X}) - \mathrm{cov}_p^{-1}\right), \quad \hat{X} \sim p$$

$$p_k^+ = p_{m_{k+1}, \Sigma_k^+}, \quad h \le 1$$

$$\rho_k^+ = \exp_{\rho_k}(-h\,\hat{g}_k) \quad \&$$

$\hat{g}_k \in T_{\rho_k} BW(\mathbb{R}^d)$ is the stochastic gradient

$$\hat{g}_k(x) = \nabla V(\hat{X}_k) + (\nabla^2 V(\hat{X}_k) - \Sigma_k^{-1})(x - m_k)$$

Question

$X_t \sim \pi_t$ $\qquad$ $Y_t \sim GP(m(\cdot), K(\cdot,\cdot))$