

Motivation

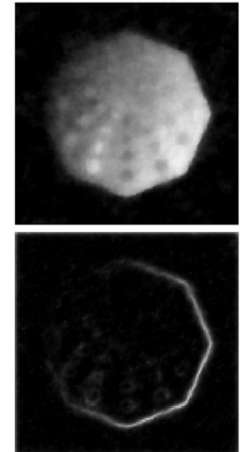
Inverse Problems and Machine Learning

High-dimensional problems in inverse problems / image reconstruction

- Reconstruct suitable solution from variational problem
- Sample posterior to quantify uncertainty (Bayes model)

$$\pi(u|f) = \frac{1}{\pi_*(f)} \pi(f|u) \pi_0(u)$$

Classical MCMC sampling (Metropolis, Gibbs ..) often inefficient, does not find modes

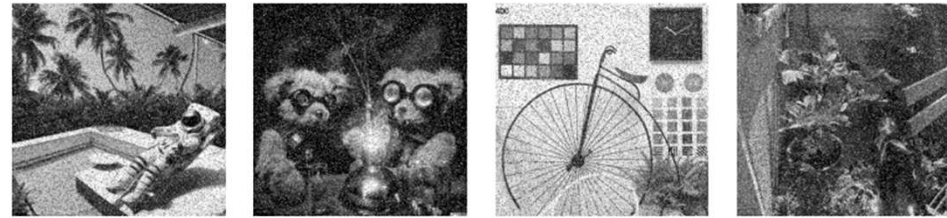


Computational Uncertainty Quantification

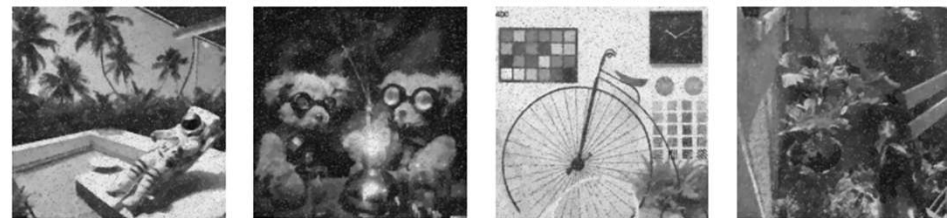
Modelling, development
of efficient computational
sampling

Here: Primal Dual Sampling,
Lorenz Kuger

Noisy image



ROF model MAP



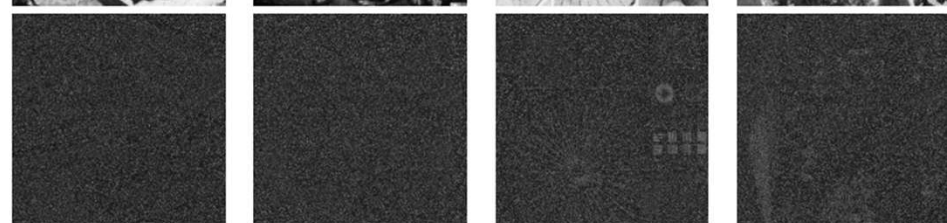
l_1 -TV MAP



l_1 -TV MMSE



l_1 -TV post. std.



Motivation

Optimization, Particles and PDEs

Simplest optimization dynamics for minimizing energy E : gradient flow

$$\frac{dx}{dt} = -\nabla E(x)$$

If initial value is considered stochastic, the distribution of x satisfies

$$\partial_t \mu = \nabla \cdot (\mu \nabla E)$$

Standard analysis of the optimization algorithm corresponds to an analysis of the Wasserstein gradient flow

$$\partial_t \mu = \nabla \cdot (\mu \nabla \mathcal{E}')$$

$$\mathcal{E}(\mu) = \int E(x) \mu(dx)$$

Motivation

Sampling, Particles and PDEs

Sampling from Gibbs energy (e.g. posterior in inverse problem)

$$\pi(dx) \sim e^{-E(x)} dx$$

Problem in classical sampling methods: computation of partition function

$$Z = \int e^{-E(x)} dx$$

Idea of score-based models: use sampling based on score

$$s = -\nabla \log \pi = \nabla E$$

Motivation

Sampling, Particles and PDEs

Langevin sampling

$$dx = -\nabla E(x) dt + \sqrt{2}dW$$

Distribution solves Fokker-Planck Equation

$$\partial_t \mu = \nabla \cdot (\mu s + \nabla \mu) = \nabla \cdot (\mu \nabla (\log \mu - \log \pi))$$

Wasserstein gradient flow for energy / relative entropy

$$\mathcal{E}(\mu) = \int E(x) \mu(dx) + \int \mu \log \mu(dx) = \int \int \mu \log \frac{\mu}{\pi}(dx)$$

Dissipation of relative entropy yields convergence to equilibrium distribution, global convergence of samples even for E not log-concave

Motivation

6

Entropy dissipation $\frac{d}{dt} RE(\mu|\mu_*) = -\mathbb{E}_\mu(|\nabla \log \frac{d\mu}{d\mu_*}|^2)$

can be combined with log-Sobolev inequality

$$\mathbb{E}_\mu(|\nabla \log \frac{d\mu}{d\mu_*}|^2) \geq C RE(\mu|\mu_*)$$

to give convergence rate (without strong assumptions on F !)

Disadvantages:

- needs Radon-Nikodym derivative (no concentration)
- not well suited for stability estimates, e.g. for inexact E
- not related to convergence analysis of gradient flow

Langevin Sampling

Sampling, Particles and PDEs

Undadjusted Langevin sampling (Doucet et al 2018)

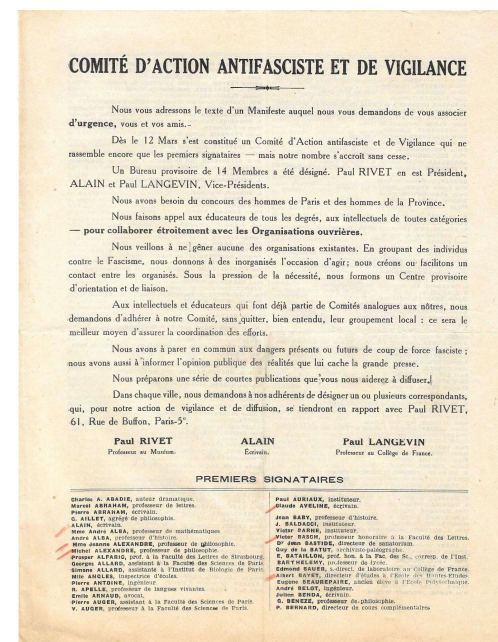
$$X^{k+1} = X^k - \tau \nabla E(X^k) + \frac{\sqrt{\tau}}{2} \omega_k, \quad \omega_k \sim N(0, I)$$

Approximate sampling of the target distribution for large k, error decreasing with time step

Convergence of discretization of SDEs, but needed uniform in time !

Various versions of Langevin methods

- Different time discretizations (e.g. forward-backward)
- Different metrics (e.g. mirror Langevin)
- Different adjustments (e.g. Metropolis adjustment)
- Different underlying dynamic (e.g. overdamped second order)



Coupling

8

A coupling between two measures is a measure on the product space with given marginals

$$\mathcal{C}(\mu, \eta) = \{\Pi \mid \Pi_{\#1} = \mu, \Pi_{\#2} = \nu\}$$

Introduced in probability to estimate distances, often total variation distance (Doebelin 1938)

Also for Wasserstein and other transport distances in diffusive processes (Chen-Li 1989, Chen-Yan 1991)

Recently used in sampling and in PDE formulations

(Perthame-Fournier 2019, Dalalyan-Karagulyan 2019, Zhang-Peyre-Fadili-Pereira 2020)

Optimal Transport Distances: Kantorovich Formulation

9

JOURNALS » EMSS » VOL. 7, NO. 1 » PP. 1–31

Transport distances for PDEs: the coupling method

Nicolas Fournier
Sorbonne Université, Paris, France

Benoît Perthame
Sorbonne Université, Paris, France

Minimization over all couplings

$$T_D(\mu, \eta) = \inf_{\Pi \in \mathcal{C}(\mu, \eta)} \mathbb{E}_{(u, v) \sim \Pi} (D(u, v))$$

$$\mathcal{C}(\mu, \eta) = \{\Pi \mid \Pi_{\#1} = \mu, \Pi_{\#2} = \nu\}$$

Wasserstein-p metric

$$W_p(\mu, \eta) = (T_D(\mu, \eta))^{1/p}, \quad D = \|u - v\|^p$$

Optimal Transport Distances for Stochastic Processes

10

Given evolution of measures μ_k, ν_k

Construction of a coupling method: construct evolution of couplings Π_k

Derive estimate

$$\mathbb{E}_{(u,v) \sim \Pi_k} (D(u, v)) \leq C_k \mathbb{E}_{(u,v) \sim \Pi_0} (D(u, v)), \quad C_k \rightarrow 0$$

Ideally take initial coupling to realize infimum

$$T_D(\mu_k, \eta_k) \leq \mathbb{E}_{(u,v) \sim \Pi_k} (D(u, v)) \leq C_k T_D(\mu_0, \eta_0)$$

Optimal Transport Distances: Kantorovich Formulation

11

Consistency with expected distance:

$$T_D(\mu, \delta_{u^*}) = \mathbb{E}_{u \sim \mu}(D(u, u^*))$$

$$\mathcal{C}(\mu, \eta) = \{\mu \otimes \delta_{u^*}\}$$

If both measures are concentrated

$$\mu = \delta_{u^*}, \nu = \delta_{v^*}$$

$$T(\mu, \nu) = D(u^*, v^*)$$

Optimal Transport Distances for Deterministic Processes

12

Concentrated measures $\mu_k = \delta_{u_k}, \nu_k = \delta_{u^*}$

Transport distance equals standard distance

$$T_D(\mu_k, \eta_k) = D(u_k, u^*)$$

Standard convergence estimate equals transport distance estimate

$$T_D(\mu_k, \eta_k) = D(u_k, u^*) \leq C_k D(u_0, u^*) = C_k T_D(\mu_0, \eta_0)$$

Optimal Transport Distances for Deterministic Processes

13

Simple example: S contractive operator

$$u_{k+1} = S(u_k), \quad v_{k+1} = S(v_k)$$

Deterministic stability

$$\|u_k - v_k\|^p \leq c^{kp} \|u_0 - v_0\|^p$$

Sampling algorithm

$$u_{k+1} = S(u_k) + \omega_k \quad \omega_k \sim Q_k$$

Coupling Analysis

14

Construct coupling via dependent random variables

$$u_{k+1} = S(u_k) + \omega_k$$

$$v_{k+1} = S(v_k) + \omega_k$$

Evolution of the joint measure in weak form

$$\mathbb{E}_{\Pi_{k+1}}(\Phi(u, v)) = \mathbb{E}_{\Pi_k} \mathbb{E}_{Q_k}(\Phi(S(u) + \omega, S(v) + \omega))$$

Note: for a test function depending on difference the additional random component drops out !

$$\mathbb{E}_{\Pi_{k+1}}(\|u - v\|^p) = \mathbb{E}_{\Pi_k}(\|S(u) - S(v)\|^p)$$

Hence, same estimate as in the deterministic setting

$$\mathbb{E}_{\Pi_{k+1}}(\|u - v\|^p) \leq c^{kp} \mathbb{E}_{\Pi_0}(\|u - v\|^p)$$

Coupling Analysis of Langevin Sampling

15

Coupling of two dependent variables (with same Wiener process W)

$$du = -\nabla F(u)dt + \sqrt{2}dW$$

$$dv = -\nabla F(v)dt + \sqrt{2}dW$$

Coupling measure satisfies the PDE

$$\begin{aligned} \partial_t \Pi = & \nabla_u \cdot (\Pi \nabla F(u)) + \nabla_v \cdot (\Pi \nabla F(v)) + \\ & \Delta_u \Pi + \Delta_v \Pi + 2 \nabla_u \cdot \nabla_v \Pi \end{aligned}$$

Simple estimate

$$\frac{d}{dt} \int \|u - v\|^2 d\Pi = -2 \int \langle \nabla F(u) - \nabla F(v), u - v \rangle d\Pi$$

For log-concave measure

$$\frac{d}{dt} \int \|u - v\|^2 d\Pi = -2\lambda \int \|u - v\|^2 d\Pi$$

Primal-dual Langevin Sampling

Sampling in large-scale image reconstruction

Energies in inverse problems often composite (data fidelity and regularization) $f(Kx) + g(x)$

For convex f , use duality to show that minimization of the composite energy equals $\min_x \max_y s(x, y)$

$$s(x, y) := g(x) + \langle Kx, y \rangle - f^*(y)$$

$$f^*(y) := \sup_x \{ \langle x, y \rangle - f(x) \}$$

Primal-dual proximal algorithms for optimization

$$\begin{aligned} Y^{n+1} &= \text{prox}_{\sigma f^*}(Y^n + \sigma K X_\theta^n) \\ X^{n+1} &= \text{prox}_{\tau g}(X^n - \tau K^T Y^n) \\ X_\theta^{n+1} &= X^{n+1} + \theta(X^{n+1} - X^n) \end{aligned}$$

Primal-dual Langevin Sampling

Sampling in large-scale image reconstruction

Unadjusted Langevin Primal-Dual algorithm proposed by Harbring et al 2023

$$\begin{aligned} Y^{n+1} &= \text{prox}_{\sigma f^*} (Y^n + \sigma K X_\theta^n) \\ \xi^{n+1} &\sim \mathcal{N}(0, I_d) \\ X^{n+1} &= \text{prox}_{\tau g} (X^n - \tau K^T Y^n) + \sqrt{2\tau} \xi^{n+1} \\ X_\theta^{n+1} &= X^{n+1} + \theta (X^{n+1} - X^n), \end{aligned}$$

Simple continuous-time model to gain understanding (mb-Ehrhardt-Kuger-Weigand 2024) $\lim_{\tau, \sigma \rightarrow 0} \frac{\sigma}{\tau} = \lambda.$

$$\begin{aligned} dX_t &\in -(\partial g(X_t) + K^* Y_t) dt + \sqrt{2} dW_t \\ dY_t &\in -\lambda(\partial f^*(Y_t) - K X_t) dt, \end{aligned}$$

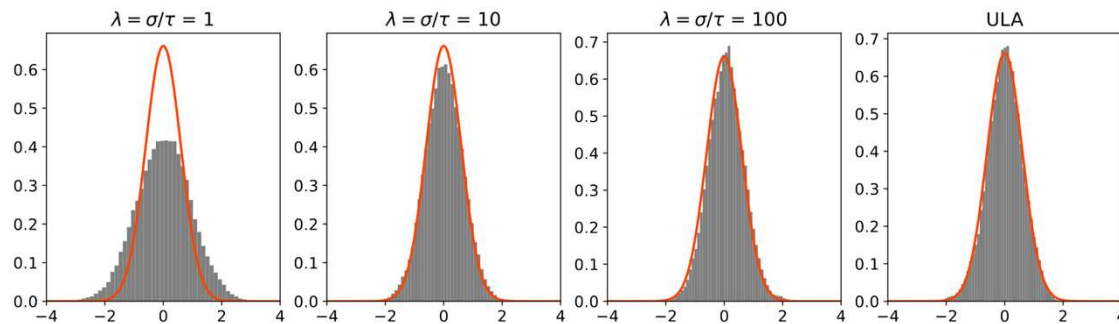
Coupling analysis can be performed with ideas from standard Langevin and the deterministic optimization

Convergence to stationary distribution, convergence of discrete algorithm to continuous formulation

Primal-dual Langevin Sampling

Consistency ?

Note: we want a special stationary distribution, concentrated in the dual variable, primal samples from correct distribution

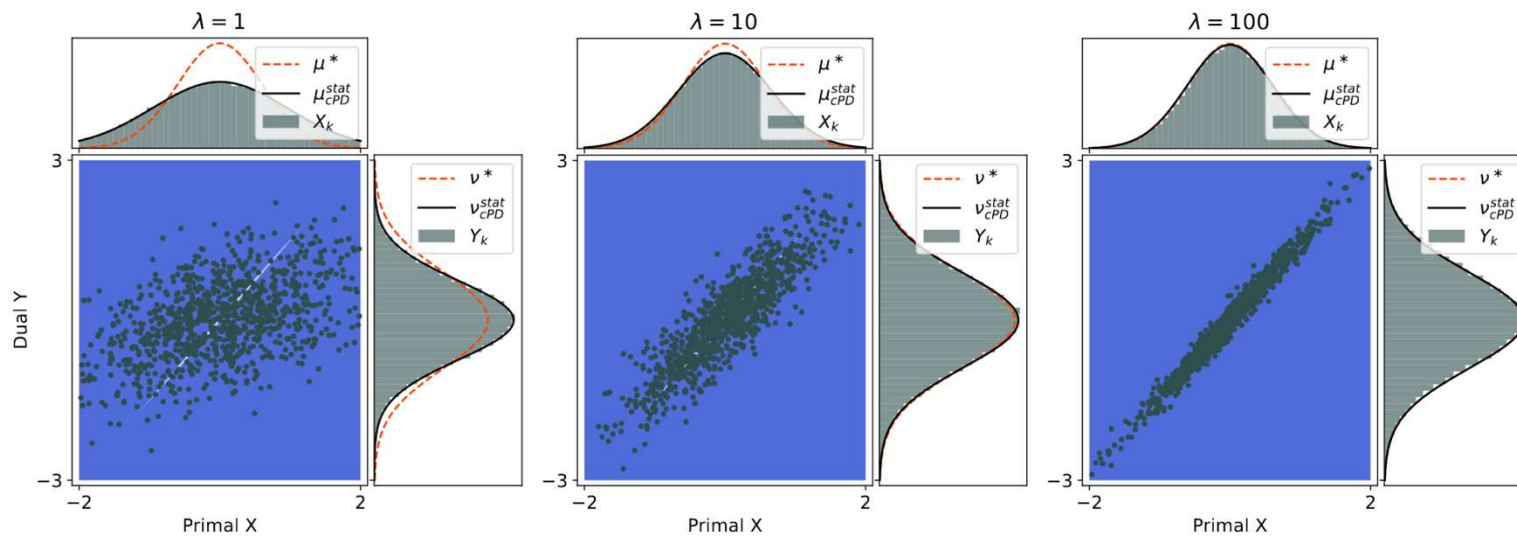


For finite I , stationary distribution is not concentrated in the dual variable:

- Explicit solutions for quadratic functionals
- Hypocoellipticity theory of the degenerate Fokker-Planck equation for smooth functionals

Primal-dual Langevin Sampling

Consistency ?



Primal-dual Langevin Sampling

Bias-corrected version

Non-concentration somehow due to Ito Formula, can be corrected with appropriate diffusion in the dual variable

$$dZ_t = -A_{\text{mod}}(Z_t) dt + B_{\text{mod}}(Z_t) dW_t$$

$$A_{\text{mod}}(z) := \begin{pmatrix} \nabla g(x) + K^T y \\ \lambda(\nabla f^*(y) - Kx) + M(x)^T(\nabla g(x) + K^T \nabla f(Kx)) \end{pmatrix}$$

$$B_{\text{mod}}(z) := \sqrt{2} \nabla(id, \nabla f \circ K)(x) = \sqrt{2} \begin{pmatrix} I \\ M(x)^T \end{pmatrix}.$$

$$M(x) := \nabla(\nabla f \circ K)^T(x)$$

Has exact stationary distribution

Disadvantage: higher-order derivatives in f !

Coupling methods

How to choose transport distance ?

Diffusion with nonconstant coefficient is expansive in standard Wasserstein metric

$$dX = \Sigma(X) dW$$

Yields

$$\partial_t \mu = \nabla \cdot (\nabla \cdot (A(x)\mu)), \quad A = \Sigma^T \Sigma$$

Canonical coupling from SDE with same Wiener process

$$\partial_t \Pi = \nabla_x \cdot \nabla_x \cdot (A(x)\Pi) + \nabla_y \cdot \nabla_y \cdot (A(y)\Pi) + \nabla_x \cdot \nabla_y \cdot ((\Sigma(x)^T \Sigma(y) + \Sigma(y)^T \Sigma(x))\Pi)$$

Open question: how to obtain non-expansiveness. Conjecture: use adapted metric (ongoing with L.Kanzler and B.Perthame)

$$d(x, y)^2 := \inf_{\xi(0)=x, \xi(1)=y} \int_0^1 \xi(t)^T A(\xi(t))^{-1} \xi(t) dt$$

Primal-dual Langevin Sampling

Asymptotic version

Use algorithm for large parameter, estimate error to correct target distribution

Construct coupling for ULPDA and ULA, analysis in transport distance for

$$d_{1,\varepsilon}(x, y, \tilde{x}) := \left(\|x - \tilde{x}\|^2 + \varepsilon \|y - \nabla f(K\tilde{x})\|^2 \right)^{1/2} \quad \varepsilon = \lambda^{-1}$$

This implies estimate for time-dependent distribution relative to exact one

$$\mathcal{W}_2(\mu_t, \mu^*) \leq \left((1 + \varepsilon \omega_{f^*} \|K\|)^{1/2} \mathcal{W}_2(\mu_0, \mu^*) + C(\varepsilon) \right) e^{-\omega_g t} + C(\varepsilon)$$

And for stationary one

$$\mathcal{W}_2(\mu_{\text{cPD}}^{\text{stat}}, \mu^*) \leq C(\varepsilon) \quad C(\varepsilon) = \sqrt{\varepsilon C_1 / \omega_g} + \mathcal{O}(\varepsilon^{3/2})$$

Score-based Diffusion Models

Generating images

Popular diffusion models can be formulated in a similar form. Given an empirical distribution of samples, evolve towards Gaussian (β used to accelerate)

$$dx = -\beta(t)X dt + \sqrt{2\beta(t)}dW$$
$$\partial_t \mu = \beta(t) \nabla \cdot (\mu x + \nabla \mu)$$

Learn score (or potential) as a neural network from (Song et al 2020)

$$\int_0^T \lambda(t) \int |s_\theta(x, t) - \nabla \log \mu|^2 \mu(dx) dt$$

To generate novel samples use Gaussian samples as starting values and solve the backward SDE

$$d\tilde{X} = \beta(T - \tilde{t})\tilde{X} d\tilde{t} + \sqrt{2\beta(T - \tilde{t})} d\tilde{W} - 2\beta(T - \tilde{t})s_\theta(\tilde{X}, T - \tilde{t})$$