Kalman-Wasserstein Gradient Flows

Franca Hoffmann

Computing and Mathematical Sciences
California Institute of Technology

Young Researchers Workshop: Ki-Net 2012-2019
CSCAMM, Oct 22, 2019

University of Maryland.
collaboration with:
Alfredo Garbuno-Inigo, Wuchen Li, Andrew Stuart (2019)
arXiv preprint, 1903.08866

accepted in:
SIAM Journal on Applied Dynamical Systems (SIADS)
The Big Picture
High-Level Overview

- Parameter calibration and uncertainty in complex computer models.
- Optimization approach and least squares.
- Bayesian approach and sampling.
- Ensemble Kalman Inversion (for optimization).
- Ensemble Kalman Sampling (for sampling).
- Gaussian Process Regression (for better sampling).
- Kalman-Wasserstein gradient flow structure.
Gradient Flow Structure

Inverse Problems

Numerics

The EKS and Mean Field Limits

Kalman-Wasserstein Space

Convergence to Equilibrium

Algorithmic Framework
Gradient Flow Structure
Optimization

Goal (Finite Dimensions)
Minimize $E : \Omega \to \mathbb{R}$, where $\Omega \subset \mathbb{R}^N$.

▶ Dynamical Formulation: find $\arg\min E(x)$ by solving

$$\dot{x}(t) = -\nabla E(x(t)), \quad x(t = 0) = x_0.$$ 

Goal (Infinite Dimensions)
Minimize $E : \mathcal{P}(\Omega) \to \mathbb{R}$, where $\mathcal{P}(X) = \{\rho \in L^1(\Omega) | \int_{\Omega} d\rho = 1\}$.

▶ Dynamical Formulation: find $\arg\min E(\rho)$ by solving

$$\partial_t \rho = -"\nabla E(\rho)", \quad \rho(t = 0) = \rho_0.$$
Quadratic Wasserstein Metric $\mathcal{W}: \mathcal{P} \times \mathcal{P} \to \mathbb{R}$

**Optimal Transport Formulation**

For $\mu, \nu \in \mathcal{P}(\Omega)$,

$$W(\mu, \nu)^2 := \inf_{\gamma} \int_{\Omega \times \Omega} |x - y|^2 d\gamma(x, y)$$

subject to $\gamma(A, \Omega) = \mu(A), \quad \gamma(\Omega, B) = \nu(B)$ for all $A, B \subset \Omega$.

**Dynamic Formulation**

For $\mu, \nu \in \mathcal{P}(\Omega)$,

$$W(\mu, \nu)^2 := \inf_{(\rho_t, \phi_t)} \int_0^1 \int_{\Omega} \langle \nabla \phi_t, \nabla \phi_t \rangle d\rho_t(x) dt$$

subject to $\partial_t \rho_t + \nabla \cdot (\rho_t \nabla \phi_t) = 0, \quad \rho_0 = \mu, \quad \rho_1 = \nu$. 
Wasserstein Gradient Flow

Tangent Vectors

- Gradient in Wasserstein metric:

$$\nabla_W E(\rho) = -\nabla \cdot \left( \rho \nabla \frac{\delta E(\rho)}{\delta \rho} \right).$$

- First variation (formal definition):

$$\lim_{\varepsilon \to 0} \left( \frac{E(\rho + \varepsilon \varphi) - E(\rho)}{\varepsilon} \right) = \int_\Omega \frac{\delta E(\rho)}{\delta \rho} [x] \varphi(x) \, dx, \quad \forall \varphi.$$  

[Ambrosio, Gigli, Savaré 2005]
Gradient Flow Structure

PDEs as GFs

Any PDE of the form

$$\partial_t \rho = \nabla \cdot (\rho \mathbf{v})$$

can be interpreted as a W-GF if there exists an energy $E : \mathcal{P}(\Omega) \to \mathbb{R}$ such that

$$\mathbf{v} = \nabla \frac{\delta E(\rho)}{\delta \rho}.$$
Kalman-Wasserstein Space

Gradient Flow Structure

\[\partial_t \rho = \nabla \cdot \left( \rho C(\rho) \nabla \frac{\delta E(\rho)}{\delta \rho} \right)\]

\[C(\rho) = \int (\theta - \overline{\theta}) \otimes (\theta - \overline{\theta}) \rho(\theta, t) d\theta, \quad \overline{\theta} = \int \theta \rho(\theta, t) d\theta.\]

- Gradient flow in \(C(\rho)\)-weighted metric.
- Suitable energy \(E\): unique attractor is the posterior for an underlying inverse problem.
- Inspired new efficient derivative-free algorithm to solve inverse problems.
Solving Inverse Problems
Inverse Problem for Parameters

Find Parameter $\theta$ from Data $y$

Let $G : \Theta \mapsto \mathcal{Y}$, and $\eta$ be noise. Then data and parameter are related by

$$y = G(\theta) + \eta, \quad \eta \sim \mathcal{N}(0, \gamma^2 I).$$

Our Setting

- Calibration and UQ for $\theta$ are both important.
- $G$ is expensive to evaluate.
- Derivatives of $G$ are not available.
Optimization Approach

Find Parameter $\theta$ from Data $y$

Let $G : \Theta \mapsto \mathcal{Y}$, and $\eta$ be noise. Then data and parameter are related by

$$y = G(\theta) + \eta, \quad \eta \sim \mathcal{N}(0, \gamma^2 I).$$

Mathematical Formulation

$$\theta^* = \arg\min_{\theta \in \Theta} \Phi_R(\theta; y),$$

$$\Phi_R(\theta; y) = \frac{1}{2\gamma^2} |y - G(\theta)|^2 + \frac{1}{2} \langle \theta, \Sigma^{-1} \theta \rangle.$$ 

Algorithms: parameter $\theta$ calibration.
Bayesian Approach

- **Prior:** $\mathbb{P}(\theta); \theta \sim \text{N}(0, \Sigma_0)$
- **Likelihood:** $\mathbb{P}(y|\theta); y - G(\theta) \sim \text{N}(0, \gamma^2 I)$
- **Posterior:** $\mathbb{P}(\theta|y); \theta|y \sim ?$

### Mathematical Formulation

$$\mathbb{P}(\theta|y) \propto \mathbb{P}(y|\theta) \times \mathbb{P}(\theta),$$

$$\mathbb{P}(\theta|y) \propto \exp\left(-\frac{1}{2\gamma^2}|y - G(\theta)|^2\right) \times \exp\left(-\frac{1}{2}\langle \theta, \Sigma_0^{-1}\theta \rangle\right)$$

$$\propto \exp\left(-\Phi_R(\theta; y)\right)$$

Algorithms: parameter $\theta$ sampling.
Parameter Calibration & UQ

\[ P(\theta|y) \propto \exp\left(-\Phi_R(\theta; y)\right) \quad \Phi_R(\theta; y) = \frac{1}{2\gamma^2}|y - G(\theta)|^2 + \frac{1}{2}\langle \theta, \Sigma_0^{-1}\theta \rangle \]

Our Setting

- Calibration: maximizing \( P(\theta|y) = \text{minimizing } \Phi_R(\theta; y) \).
- UQ: sample from \( P(\theta|y) \).
- \( G \) is expensive to evaluate.
- Derivatives of \( G \) are not available.

Goals

- Methods to generate approximate samples from posterior.
- Mathematical framework for analysis using gradient flow structure.
Ensemble Kalman Inversion (EKI)

Continuous Time Formulation

\[ \dot{\theta}(j) = -\frac{1}{J} \sum_{k=1}^{J} \frac{1}{\gamma^2} \left\langle G(\theta(k)) - \bar{G}, G(\theta(j)) - y \right\rangle \left( \theta(k) - \bar{\theta} \right), \]

\[ \bar{\theta} = \frac{1}{J} \sum_{k=1}^{J} \theta(k), \quad \bar{G} = \frac{1}{J} \sum_{k=1}^{J} G(\theta(k)). \]

[Evensen 1994], [Inglesias, Law, Stuart 2013], [Ernst, Sprungk, Starkloff 2015], [Schillings, Stuart 2017], [Herty, Visconti 2019]

▶ Sample from prior instead of posterior.
▶ Evolve particle ensemble: derivative-free algorithm.
▶ Tool for optimization: collapses to max \( \mathbb{P}(\theta | y) \).


Ensemble Kalman Sampling (EKS)

\[ C(\theta) = \frac{1}{J} \sum_{k=1}^{J} (\theta^{(k)} - \bar{\theta}) \otimes (\theta^{(k)} - \bar{\theta}). \]

**Continuous Time Formulation**

\[
\dot{\theta}^{(j)} = -\frac{1}{J} \sum_{k=1}^{J} \frac{1}{\gamma^2} \left\langle G(\theta^{(k)}) - \bar{G}, G(\theta^{(j)}) - y \right\rangle \left( \theta^{(k)} - \bar{\theta} \right)
- C(\theta)\Sigma^{-1}_0 \theta^{(j)} + \sqrt{2C(\theta)} \dot{W}^{(j)}
\]

[Garbuno-Inigo, Li, H., Stuart 2019 (preprint)]

- Add damping term related to prior [Chada, Stuart, Tong 2019 (preprint)].
- Perturb particles instead of data [Kovachki, Stuart 2019].
- New noise covariance structure.
Numerics:
EKI vs EKS vs MCMC
Example: Elliptic BVP

- 1D problem for $x \in [0, 1]$,

$$-\frac{d}{dx} \left( \exp(\theta_1) \frac{d}{dx} p(x) \right) = 1,$$

with boundary conditions $p(0) = 0$ and $p(1) = \theta_2$.

- Explicit solution is available and we define

$$G(\theta) = \begin{pmatrix} p(x_1) \\ p(x_2) \end{pmatrix}.$$

- Run EKI with $J = 1000$. 

Figure: Contour plots: $\Phi(\theta) = \frac{1}{2} \gamma^2 \left| y - G(\theta) \right|^2$. 
Example: Elliptic BVP

- 1D problem for $x \in [0, 1]$,

$$\frac{d}{dx} \left( \exp(\theta_1) \frac{d}{dx} p(x) \right) = 1,$$

with boundary conditions $p(0) = 0$ and $p(1) = \theta_2$.

- Explicit solution is available and we define

$$G(\theta) = \begin{pmatrix} p(x_1) \\ p(x_2) \end{pmatrix}.$$

- Run EKI with $J = 1000$.

Figure: Contour plots: $\Phi(\theta) = \frac{1}{2\gamma^2} | y - G(\theta) |^2$. 
Example: Elliptic BVP

- 1D problem for $x \in [0, 1]$, 

  $$ - \frac{d}{dx} \left( \exp(\theta_1) \frac{d}{dx} p(x) \right) = 1, $$

  with boundary conditions $p(0) = 0$ and $p(1) = \theta_2$.

- Explicit solution is available and we define

  $$ G(\theta) = \begin{pmatrix} p(x_1) \\ p(x_2) \end{pmatrix}. $$

- Run EKS with $J = 1000$.

- Compare with exact MCMC.
Example: Elliptic BVP

- 1D problem for $x \in [0, 1]$,

$$- \frac{d}{dx} \left( \exp(\theta_1) \frac{d}{dx} p(x) \right) = 1,$$

with boundary conditions $p(0) = 0$ and $p(1) = \theta_2$.

- Explicit solution is available and we define

$$G(\theta) = \begin{pmatrix} p(x_1) \\ p(x_2) \end{pmatrix}.$$

- Run EKS with $J = 1000$.

- Compare with exact MCMC.
Example: Elliptic BVP

- 1D problem for \( x \in [0, 1] \),

\[
- \frac{d}{dx} \left( \exp(\theta_1) \frac{d}{dx} p(x) \right) = 1,
\]

with boundary conditions \( p(0) = 0 \) and \( p(1) = \theta_2 \).

- Explicit solution is available and we define

\[
G(\theta) = \begin{pmatrix} p(x_1) \\ p(x_2) \end{pmatrix}.
\]

- Run EKS with \( J = 1000 \).
- Compare with exact MCMC.
From EKS to Kalman-Wasserstein Gradient Flow
EKS: Approximation 1

\[ \dot{\theta}^{(j)} = -\frac{1}{J} \sum_{k=1}^{J} \frac{1}{\gamma^2} \left\langle G(\theta^{(k)}) - \bar{G}, G(\theta^{(j)}) - y \right\rangle \left( \theta^{(k)} - \bar{\theta} \right) - C(\theta) \Sigma_0^{-1} \theta^{(j)} + \sqrt{2C(\theta)} \dot{W}^{(j)}, \]

\[
C(\theta) = \frac{1}{J} \sum_{k=1}^{J} \left( \theta^{(k)} - \bar{\theta} \right) \otimes \left( \theta^{(k)} - \bar{\theta} \right). 
\]

Linear Approximation

\[
(G(\theta^{(k)}) - \bar{G}) \approx A(\theta^{(k)} - \bar{\theta}), \quad A := DG(\theta^{(j)}).
\]
EKS: Approximation 1

Preconditioned Langevin Equation

\[ \dot{\theta}^{(j)} = -C(\theta) \nabla \Phi_{R}(\theta^{(j)}) + \sqrt{2C(\theta)} \dot{W}^{(j)}. \]

\[ \Phi_{R}(\theta) \approx \frac{1}{2} \gamma^2 |y - A\theta|^2 + \frac{1}{2} |\Sigma_{0}^{-\frac{1}{2}} \theta|^2. \]

- If G linear, this is EKS.
- If G non-linear, expect solution to be close to EKS solution if particles close together (conjecture!).
- Preconditioning accelerates [Matthews, Leimkuhler, Weare 2018].
EKS: Approximation 2

\[ C(\rho) := \int (\theta - \bar{\theta}) \otimes (\theta - \bar{\theta}) \rho(\theta, t) d\theta, \quad \bar{\theta} := \int \theta \rho(\theta, t) d\theta. \]

**Mean Field Limit**

\[ \dot{\theta} = -C(\rho) \nabla \Phi_R(\theta) + \sqrt{2C(\rho)} \dot{W}. \]

**Nonlinear Fokker-Planck Equation**

\[ \partial_t \rho = \nabla \cdot (\rho C(\rho) \nabla \Phi_R) + C(\rho) : D^2 \rho, \]

[Li, Din 2019 (preprint)]
Connection To Bayesian Inversion

\[ \partial_t \rho = \nabla \cdot (\rho C(\rho) \nabla (\Phi_R + \log \rho)) , \quad E(\rho) = \int \rho (\Phi_R + \log \rho) \, d\theta . \]

Manifold of Stationary States

\[ \rho(\theta) = \delta_v(\theta) \text{ for some } v \in \mathbb{R}^d \iff C(\rho) = 0 . \]

Steady State

Equilibrium solution to non-linear Fokker-Planck equation:

\[ \rho_\infty(\theta) := \frac{e^{-\Phi_R(\theta)}}{\int e^{-\Phi_R(\theta)} \, d\theta} . \]

This is the density of \( \mathbb{P}(\theta|y) \).
Kalman-Wasserstein Space
Gradient Flow Structure

For $\Omega \subseteq \mathbb{R}^d$ convex set, define $\mathcal{P}_+ := \{\rho \in \mathcal{P} : \rho > 0 \text{ a.e.}, \rho \in C^\infty(\Omega)\}$.

Kalman-Wasserstein Metric $\mathcal{W}_C : \mathcal{P}_+ \times \mathcal{P}_+ \to \mathbb{R}$

For $\mu, \nu \in \mathcal{P}_+$,

$$\mathcal{W}_C(\mu, \nu)^2 := \inf_{(\rho_t, \phi_t)} \int_0^1 \int_\Omega \langle \nabla \phi_t, C(\rho_t) \nabla \phi_t \rangle \rho_t \, dx$$

subject to \quad $\partial_t \rho_t + \nabla \cdot (\rho_t C(\rho_t) \nabla \phi_t) = 0, \rho_0 = \mu, \rho_1 = \nu$,

Theorem

Given a finite functional $E : \mathcal{P}_+ \to \mathbb{R}$, the gradient flow of $E(\rho)$ in $(\mathcal{P}_+, \mathcal{W}_C)$ satisfies

$$\partial_t \rho = \nabla \cdot \left( \rho C(\rho) \nabla \frac{\delta E}{\delta \rho} \right).$$

[Garbuno-Inigo, Li, H., Stuart 2019 (preprint)]
**Energy:** Kullback-Leibler divergence

\[
E(\rho) = \int (\Phi_R + \ln \rho(t)) \rho(t) \, d\theta
\]

\[
= \int \frac{\rho(t)}{\rho_\infty} \ln \left( \frac{\rho(t)}{\rho_\infty} \right) \rho_\infty \, d\theta + \ln \left( \int e^{-\Phi_R(\theta)} \, d\theta \right)
\]

\[
= \text{KL}(\rho(t)\|\rho_\infty) + c
\]

**Euler-Lagrange condition:**

\[
\frac{\delta E}{\delta \rho} = \Phi_R(\theta) + \ln \rho(\theta) = c \quad \text{on supp } (\rho)
\]

**Unique solution:** posterior

\[
\rho_\infty(\theta) := \frac{e^{-\Phi_R(\theta)}}{\int e^{-\Phi_R(\theta)} \, d\theta}.
\]
Energy dissipation:

\[ \frac{d}{dt} E(\rho) = - \int \rho \left| C(\rho)^{\frac{1}{2}} \nabla (\Phi_R + \ln \rho) \right|^2 \, d\theta. \]

Hence \( E \searrow \) along paths until \( C(\rho) = 0 \) or \( \rho = \rho_\infty \).

Fisher-Information: for any covariance matrix \( \Lambda \),

\[ \mathcal{I}_\Lambda(\rho\|\rho_\infty) := \int \rho \left\langle \nabla \ln \left( \frac{\rho}{\rho_\infty} \right) , \, \Lambda \nabla \ln \left( \frac{\rho}{\rho_\infty} \right) \right\rangle \, d\theta. \]

Kalman-Fisher information: For \( \Lambda = C(\rho) \),

\[ \frac{d}{dt} \text{KL}(\rho(t)\|\rho_\infty) = -\mathcal{I}_C(\rho(t)\|\rho_\infty). \]
Convergence to Equilibrium
Nonlinear forward map

Theorem (Decay to Equilibrium)

Assume $\alpha > 0$ and $\lambda > 0$ exists such that

$$C(\rho(t)) \geq \alpha I_d, \quad D^2 \Phi_R \geq \lambda I_d.$$ 

If $\text{KL}(\rho_0 \| \rho_\infty) < \infty$ then there is $c > 0$ such that

$$\|\rho(t) - \rho_\infty\|_{L^1(\mathbb{R}^d)} \leq ce^{-\alpha \lambda t}.$$ 

[Garbuno-Inigo, Li, H., Stuart 2019 (preprint)]
Proof

\( D^2 \Phi_R \geq \lambda I_d \) guarantees log Sobolev inequality [Bakry, Émery 1985]:

\[
KL(\rho(t)\|\rho_\infty) \leq \frac{1}{2\lambda} I_d(\rho(t)\|\rho_\infty) \quad \forall \rho.
\]

\( C(\rho(t)) \geq \alpha I_d \) gives

\[
\frac{d}{dt} KL(\rho(t)\|\rho_\infty) = -C(\rho(t)\|\rho_\infty)
\leq -\alpha I_d(\rho(t)\|\rho_\infty) \leq -2\alpha \lambda KL(\rho(t)\|\rho_\infty).
\]

By Csiszár-Kullback inequality:

\[
\frac{1}{2} \|\rho(t) - \rho_\infty\|_{L^1(\mathbb{R}^d)}^2 \leq KL(\rho(t)\|\rho_\infty) \leq KL(\rho_0\|\rho_\infty) e^{-2\alpha \lambda t}.
\]
Linear forward map $G(\theta) = A\theta$

Theorem (Linear Inverse Problem).

- Closed equations for the moments.
- Mean field limit of EKS: Gaussians remain Gaussians.
- Gaussians converge exponentially fast to $\rho_\infty$ in $L^1(\mathbb{R}^d)$ as $t \to \infty$.

[Garbuno-Inigo, Li, H., Stuart 2019 (preprint)]

Theorem (Decay to Equilibrium)

If $\rho^1_t, \rho^2_t$ solutions to KW-GF with initial conditions $\rho^1_0, \rho^2_0$ respectively, then under suitable conditions on $\rho^1_0, \rho^2_0, A, \Sigma_0, \gamma$,

$$W(\rho^1_t, \rho^2_t) \leq Ce^{-t}W(\rho^1_0, \rho^2_0).$$

where $C$ only depends on the first two moments of $\rho^1_0, \rho^2_0$, and on $A, \Sigma_0$.

[Carrillo, Vaes 2019 (preprint)]
Calibrate - Emulate - Sample

Cleary, Garbuno, Lan, Schneider, Stuart (2019)
arXiv preprint, 1911.+++++

Calibrate
\[ y = G(\theta) + \eta \]

Emulate
\[ G^{(M)}(\theta) \approx G(\theta) \]

Sample
\[ y = G^{(M)}(\theta) + \eta \]
Gaussian Process Accelerated Sampling

- From EKS we generate approximate posterior samples \( \{ \theta^{(i)}, G(\theta^{(i)}) \}_{i=1}^J \).
- Use parameter-output pairs to train a Gaussian Process (GP) emulator \( G_J(\cdot) \).

- Define \( \Phi_J(\theta; y) = \frac{1}{2\gamma^2} |y - G_J(\theta)|^2 \). Evaluation of \( \Phi_J \) is fast.
- Sample approximate posterior

\[
P_J(\theta|y) \propto \exp\left(-\Phi_J(\theta; y)\right) \times \exp\left(-\frac{1}{2} \langle \theta, \Sigma_0^{-1} \theta \rangle\right).
\]
Conclusions

- New algorithm to generate approximate posterior samples for inverse problems;
- Introduced Kalman-Wasserstein space;
- New algorithmic framework Calibrate-Emulate-Sample.

Open Questions (Theory & Practice):

- Applications of the Calibrate-Emulate-Sample framework.
- How close are dynamics of EKS to KW flow for non-linear G?
- Convergence to equilibrium for non-linear G.
- Properties of the Kalman-Wasserstein space and related functional inequalities.
- General matrix $K(\rho)$: optimal rates of convergence?
References


