A Wasserstein perspective on inference

Luca Raffo
EPFL, Institute of Mathematics - luca.raffo@epfl.ch
Capital Fund Management - luca.raffo@cfm.com

June 2025

Contents

1 Introduction

Score matching, introduced in [4], yields a new heuristic to estimate continuous statistical models where the probability density function is known only up to a multiplicative normalization constant. The method is shown to be locally consistent under identifiability of the model, and the estimation does not require to compute the normalization constant.

Here we propose a new point of view that shows that score matching is equivalent to searching for the model that minimizes the Wasserstein gradient (under empirical expectation) of the KL divergence between the real density and the estimated one.

Then we move on to show a deep connection between score matching and minimum probability flow. The latter, introduced in [6], is a different method for estimating statistical models, developed for discrete domains such as Ising models.

This framework can be generalized to different functionals and may lead to new methods for parametric statistical inference. We show and example with MMD divergences.

2 Background on Score Matching

As in usual statistical inference frameworks, we start from a given set of datapoints $D = \{x^{(1)}, ..., x^{(n)}\}$ with $x^{(i)} \in \mathbb{R}^d$, sampled via $X^{(1)}, ..., X^{(n)} \overset{\text{iid}}{\sim} \mu_{true}$, where μ_{true} is the real (unknown) distribution whose density with respect to the Lebesgue measure is f_{true}.
We want to model the distribution with π_θ, absolutely continuous with respect to Lebesgue, and with density $f_\theta(x) = \frac{1}{Z_\theta} e^{-H_\theta(x)}$.

Score matching reads

$$\begin{aligned}\theta_{SM} &= \arg\min_\theta \mathbb{E}_{\mu_{true}}[\|\nabla_x \log f_\theta - \nabla_x \log f_{true}\|_2^2] \\ &= \arg\min_\theta \mathbb{E}_{\mu_{true}}[(\nabla_x H_\theta)^2 - 2\Delta_x H_\theta] \text{ after integrating by parts},\end{aligned}$$

whose sample version is

$$\theta^*_{SM} = \arg\min_\theta \mathbb{E}_D[(\nabla_x H_\theta)^2 - 2\Delta_x H_\theta],$$

which does not require the normalization constant Z_θ and can be computed from the data.

In [4] it is shown that θ_{SM} is locally consistent if the model is identifiable (i.e. under the condition that if $\theta_1 \neq \theta_2$ then $\pi_{\theta_1} \neq \pi_{\theta_2}$), and the sample version is asymptotically equivalent to the population one due to the strong law of large numbers.

3 Background on Wasserstein Gradient Flows

The main objective of this section is to unify the notation regarding flows of measures and to define properly Wasserstein gradient flows.

3.1 Flows of measures

Let us sample $X_0 \sim \mu_0$, with $d\mu_0 = f_0 \, d\lambda$ and let $v_t : \mathbb{R}^d \to \mathbb{R}$ be any vector field. If we evolve our particle via

$$\dot{X}_t = v_t(X_t), \tag{1}$$

we find out that $\mu_t := Law(X_t)$ is absolutely continuous with respect to Lebesgue, it has finite second moment, and its density satisfies the continuity equation, i.e. it satisfies

$$\partial_t f_t + \nabla \cdot (v_t f_t) = 0 \tag{2}$$

in weak sense. The proof can be found in [1].

3.2 Background on Wasserstein spaces

Consult manuscript I.

3.3 Wasserstein gradient flows

Given a functional $\mathcal{F} : \mathcal{P}_2^{ac}(\lambda) \to \mathbb{R}$ with bounded first variation, we define its Wasserstein gradient at $\mu \in \mathcal{P}_2^{ac}(\lambda)$ as

$$\begin{aligned} \nabla_{\mathcal{W}_2} \mathcal{F}(\mu) : \mathbb{R}^d &\to \mathbb{R}^d \\ x &\mapsto \nabla[\delta \mathcal{F}(\mu)](x). \end{aligned}$$

Now we fix a functional $\mathcal{F}$ with bounded first variation, and use $-\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)$ as our vector field v_t in (1), so that μ_t will evolve via

$$\partial_t f_t + \nabla \cdot (-\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t) f_t) = 0, \tag{3}$$

that is known as *Wasserstein gradient flow* of μ_t with respect to $\mathcal{F}$, started at μ_0.

There are many vector fields v_t such that the ODE (1) induces the PDE (3) (this is due to the fact that the divergence $\nabla\cdot$ has a *big kernel*, so that if v_t satisfies (3), then $v_t + \eta_t$ satisfies it too, where η_t is any divergence free vector field), it turns out that the *most economical* one, i.e. the one which minimizes $\|v_t\|^2_{L_2(\mu_t)}$ is

$$v_t = -\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t). \tag{4}$$

Again, we suggest consulting [1] for a proof.

An important property about Wasserstein gradient flows, in similar analogy with gradient flows in finite dimensional spaces, is that if μ_t is following the gradient flow with respect fo $\mathcal{F}$, then

$$\begin{aligned} \frac{d}{dt}\mathcal{F}(\mu_t) &= -\langle \nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t), \nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\rangle_{L_2(\mu_t)} && (5)\\ &= -\|\nabla_{\mathcal{W}_2}\mathcal{F}(\mu_t)\|_{L_2(\mu_t)}, && (6) \end{aligned}$$

showing that these gradient flows dissipate energy along the flow, providing a principled approach to minimizing functionals.

Clearly if a functional $\mathcal{F}$ is strongly displacement convex, then it has a unique minima μ^* (to see this, if there were two such minima, we can arrive to a contradiction by considering the geodesic interpolating between the two of them).

The main result for this section, denoted in [2] as Poljak–Łojasiewicz inequality, states that the Wasserstein gradient flow with respect to a strongly displacement convex $\mathcal{F}$, started at any $\mu_0 \in \mathcal{P}_2^{ac}(\lambda)$, converges exponentially fast towards the unique minimizer $\mu^* \in \mathcal{P}_2^{ac}(\lambda)$.

It turns out that $\mathcal{F}(\cdot) = \mathcal{D}_{KL}(\cdot\|\pi_\theta)$ is (strongly) displacement convex, so it is reasonable trying to minimize it via Wasserstein gradient flows.

4 Score Matching and Wasserstein Gradient Flows

Let us come back to our usual framework of μ_{true} and π_θ. The idea is that if $\mu_{true} \approx \pi_\theta$, then

$$\mathcal{D}_{KL}(\mu_{true} \| \pi_\theta) \approx 0, \tag{7}$$

so that the Wasserstein gradient flow of $\mathcal{D}_{KL}(\cdot\|\pi_\theta)$ (which is playing the role of $\mathcal{F}(\cdot)$) started from $\mu_0 := \mu_{true}$ will be almost stationary. For a sample $X_0 \sim \mu_{true}$, since $\dot{X}_0 = -\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot||\pi_\theta)|_{\mu_0}(X_0)$ is the starting point of the *most economical* ODE inducing (3), the requirement (7) naturally translates into finding

$$\theta^* := \arg\min_\theta \mathbb{E}_{\mu_0}[\| - \nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot\|\pi_\theta)|_{\mu_0}\|_2^2]. \tag{8}$$

This can also be derived from (5) and (6), asking for $|\frac{d}{dt}\mathcal{D}_{KL}(\cdot\|\pi_\theta)|_{\mu_0}| \approx 0$.

We know that $\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot\|\pi_\theta)|_{\mu_0} = \nabla_x H_\theta + \nabla_x \log f_{true}$ (consult manuscript I), so that (8) is aiming to find

$$\begin{aligned}\theta^* &= \arg\min_\theta \mathbb{E}_{\mu_{true}}[(-\nabla_x H_\theta - \nabla_x \log f_{true})^2] \\ &= \arg\min_\theta \mathbb{E}_{\mu_{true}}[(\nabla_x \log f_\theta - \nabla_x \log f_{true})^2]\end{aligned}$$

that is precisely the starting point of score matching.

On a more technical note, in [2] they show that from the gradient flow perspective, the more H_θ is convex, the faster the convergence rate of μ_0 towards π_θ; we suggest it would be worth investigating whether the more convex H_θ is, the faster the score matching algorithm converges.

To show another interesting connection, notice that the Wasserstein gradient flow of $\mathcal{D}_{KL}(\cdot\|\pi_\theta)$ is the Fokker–Planck equation with potential H_θ, i.e.

$$\partial_t f_t + \nabla \cdot (-\nabla (\log f_t + H_\theta) f_t) = 0 \quad \iff \quad \partial_t f_t = \Delta f_t + \nabla \cdot (f_t \nabla H_\theta),$$

and therefore score matching is in turn equivalent to minimizing the empirical version of $|\frac{d}{dt}\mathcal{D}_{KL}(\mu_t\|\pi_\theta)|_{t=0}|$, as μ_t satisfies the Fokker Planck equation with potential H_θ, and $\mu_0 := \mu_{true}$.

5 Score Matching and Minimum Probability Flow

Minimum probability flow, introduced in [6], is a method for statistical inference on discrete models.

The idea is the following: we have samples $D = \{x^{(1)}, ..., x^{(n)}\}$ with $x^{(i)}$ from μ_{true} unknown, and we want to infer our best approximation of μ_{true} inside the family $(\pi_\theta)_{\theta\in\Theta}$, given the datapoints D. To do so, we fix a Markov kernel K_θ that leaves π_θ invariant, that is,

$$K_\theta^\top \pi_\theta = 0.$$

We then consider the probability flow starting from $\mu_0 := \mu_{true}$, evolving according to

$$\partial_t \mu_t = K_\theta^\top \mu_t. \tag{9}$$

The MPF objective measures how fast μ_t initially departs from equilibrium under this dynamics. Specifically, it minimizes the initial rate of change of the Kullback–Leibler divergence between μ_t and π_θ:

$$\theta_{MPF} = |\frac{d}{dt}\mathcal{D}_{\text{KL}}(\mu_t \| \pi_\theta)|_{t=0}|.$$

Intuitively, this quantity measures the instantaneous "probability flow" from the data distribution toward regions where the model π_θ assigns probability mass. The optimal parameters θ_{MPF} are those for which the empirical distribution μ_{true} is already approximately stationary under the dynamics defined by $K_{\theta_{MPF}}$, hence exhibiting minimal probability flow. This criterion can be expressed smoothly as

$$\theta_{MPF} = \sum_{x\in D}\sum_{x'\notin D} K_\theta(x', x),$$

corresponding to the total outflow of probability from data states to non-data states.

This perspective makes evident the analogy with score matching in the continuous setting: both approaches minimize the initial time derivative of the Kullback–Leibler divergence along a dynamics that relaxes toward π_θ, the

Fokker–Planck flow in the continuous case, and a Markov chain flow in the discrete one.

Remarkably, in [5] they showed that there exists a generalization of the Wasserstein metric on discrete spaces dependent on the choice of K_θ (remember that a priori there exist many Markov kernel which are in equilibrium with π_θ), such that the flow (9) if the Wasserstein gradient flow of the discrete KL divergence. Here speed of convergence of the flow is more subtle, as the flow depends on the choice of the kernel. In [3] they investigate Ricci curvature of Markov chains, and this could be a good starting point for the study of convergence of the flows, and for the study of the speed of convergence of minimum probability flow.

6 Towards MMD

The reasoning developed so far for score matching and minimum probability flow can be generalized to other statistical discrepancies between probability measures. In particular, we consider the case of the *Maximum Mean Discrepancy* (MMD), a widely used distance in kernel-based inference and generative modeling.

For a detailed introduction to MMD distances and their reproducing-kernel interpretation, we refer to the companion note manuscript II.

Given two measures μ and ν on $\mathbb{R}^d$, and a positive-definite kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with reproducing kernel Hilbert space (RKHS) $\mathcal{H}_k$, the squared Maximum Mean Discrepancy is defined as

$$\mathrm{MMD}_k^2(\mu,\nu) := \|\mathbb{E}_{x\sim\mu}[k(x,\cdot)] - \mathbb{E}_{y\sim\nu}[k(y,\cdot)]\|_{\mathcal{H}_k}^2. \tag{10}$$

It can be expanded as

$$\mathrm{MMD}_k^2(\mu,\nu) = \mathbb{E}_{x,x'\sim\mu}[k(x,x')] + \mathbb{E}_{y,y'\sim\nu}[k(y,y')] - 2\,\mathbb{E}_{x\sim\mu,y\sim\nu}[k(x,y)].$$

This quantity vanishes if and only if $\mu = \nu$, provided some regularity conditions regarding k.

Our goal is to define a new inference principle analogous to score matching, by replacing the KL divergence with the MMD functional. Specifically, we want to find the model π_θ such that the Wasserstein gradient of $\mathrm{MMD}_k^2(\cdot, \pi_\theta)$, evaluated at μ_{true}, is as small as possible.

Our strategy, which involves minimizing the Wasserstein norm of the Wasserstein gradient of the functional, leads to

$$\theta_k^* = \arg\min_\theta \mathbb{E}_{x\sim\mu_{true}} \left[\left\| \nabla_{\mathcal{W}_2} \mathrm{MMD}_k^2(\cdot, \pi_\theta)\big|_{\mu_{true}}(x) \right\|_2^2 \right]. \tag{11}$$

Intuitively, this corresponds to requiring that μ_{true} is almost stationary under the Wasserstein gradient flow of the MMD functional towards π_θ, in perfect analogy with what we did for the KL divergence in the score-matching framework.

Our method requires to compute the Wasserstein gradient of MMD_k^2. Taking the first variation with respect to μ, we find

$$\delta\,\mathrm{MMD}_k^2(x) = 2\Big[(k*\mu)(x) - (k*\pi_\theta)(x)\Big], \qquad (k*\nu)(x) := \int_{\mathbb{R}^d} k(x,z)\,d\nu(z).$$

Hence, the Wasserstein gradient is obtained by differentiating with respect to x:

$$\nabla_{\mathcal{W}_2}\mathrm{MMD}_k^2(\mu, \pi_\theta)(x) = \nabla_x\big[\delta\,\mathrm{MMD}_k^2(x)\big] = 2\Big[\nabla_x(k*\mu)(x) - \nabla_x(k*\pi_\theta)(x)\Big]. \tag{12}$$

By exchanging the derivative and the integral, we obtain

$$\nabla_x(k*\nu)(x) = \int_{\mathbb{R}^d} \nabla_x k(x,z)\,d\nu(z),$$

so that

$$\nabla_{\mathcal{W}_2}\mathrm{MMD}_k^2(\mu, \pi_\theta)(x) = 2\left[\int_{\mathbb{R}^d} \nabla_x k(x,z)\,d\mu(z) - \int_{\mathbb{R}^d} \nabla_x k(x,y)\,d\pi_\theta(y)\right]. \tag{13}$$

Evaluating at $\mu = \mu_{true}$ gives

$$\nabla_{\mathcal{W}_2}\mathrm{MMD}_k^2(\cdot, \pi_\theta)\big|_{\mu_{true}}(x) = 2\left[\mathbb{E}_{Z\sim\mu_{true}}[\nabla_x k(x,Z)] - \mathbb{E}_{Y\sim\pi_\theta}[\nabla_x k(x,Y)]\right]. \tag{14}$$

We now investigate explicit forms for common kernels.

1. *Linear kernel:*

   $$k(x,y) = x^\top y, \qquad \nabla_x k(x,y) = y.$$

   Therefore,

   $$\nabla_{\mathcal{W}_2} \mathrm{MMD}_k^2\big|_{\mu_{true}}(x) = 2\Big[\mathbb{E}_{Z\sim\mu_{true}}[Z] - \mathbb{E}_{Y\sim\pi_\theta}[Y]\Big],$$

   which is constant in x, and the loss reduces to the squared difference between the means:

   $$\mathcal{L}_{\mathrm{lin}}(\theta) = 4\,\big\|\mathbb{E}_{Z\sim\mu_{true}}[Z] - \mathbb{E}_{Y\sim\pi_\theta}[Y]\big\|_2^2.$$

   Hence, the linear kernel version simply enforces mean matching between μ_{true} and π_θ.

2. *Polynomial kernel:*

   $$k(x,y) = (x^\top y + c)^p, \qquad \nabla_x k(x,y) = p\,(x^\top y + c)^{p-1} y.$$

   The gradient becomes

   $$\nabla_{\mathcal{W}_2} \mathrm{MMD}_k^2\big|_{\mu_{true}}(x) = 2p\Big[\mathbb{E}_{Z\sim\mu_{true}}[\,y\,(x^\top Z+c)^{p-1}\,]-\mathbb{E}_{Y\sim\pi_\theta}[\,Y\,(x^\top Y+c)^{p-1}\,]\Big],$$

   and the corresponding loss reads

   $$\mathcal{L}_{\mathrm{poly}}(\theta) = 4p^2\,\mathbb{E}_{x\sim\mu_{true}}\Big\|\mathbb{E}_{Z\sim\mu_{true}}[\,Z\,(x^\top Z+c)^{p-1}\,]-\mathbb{E}_{Y\sim\pi_\theta}[\,Y\,(x^\top Y+c)^{p-1}\,]\Big\|_2^2.$$

3. *Gaussian RBF kernel:*

   $$k(x,y) = \exp\!\left(-\frac{\|x-y\|^2}{2\sigma^2}\right), \qquad \nabla_x k(x,y) = \frac{y-x}{\sigma^2}\,k(x,y).$$

   In this case,

   $$\nabla_{\mathcal{W}_2} \mathrm{MMD}_k^2\big|_{\mu_{true}}(x) = \frac{2}{\sigma^2}\,\big[\mathbb{E}_{Z\sim\mu_{true}}\big[(Z-x)\,k(x,Z)\big] - \mathbb{E}_{Y\sim\pi_\theta}\big[(Y-x)\,k(x,Y)\big]\big],$$

   and the loss takes the form

   $$\mathcal{L}_{\mathrm{RBF}}(\theta) = \frac{4}{\sigma^4}\,\mathbb{E}_{x\sim\mu_{true}}\Big\|\mathbb{E}_{Z\sim\mu_{true}}[(Z-x)\,k(x,Z)]-\mathbb{E}_{Y\sim\pi_\theta}[(Y-x)\,k(x,Y)]\Big\|_2^2.$$

Each choice of kernel therefore yields a different “moment-matching” condition between μ_{true} and π_θ, ranging from simple mean matching (linear kernel) to high-order nonlinear dependencies (polynomial and Gaussian kernels).

7 Numerical Experiments

We implement our methods to infer the covariance matrix of Gaussian model.

As we saw, minimizing the Wasserstein gradient of the KL divergence led to the score matching loss, while for the MMD we derived three loss functions. The MMD loss with linear kernel will not be informative on the covariance, as it only looks at the mean of the models.

The implementation and results can be found here.

References

[1] L. Ambrosio, N. Gigli, and G. Savare. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Lectures in Mathematics. ETH Zürich. Birkhäuser Basel, 2008.

[2] Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, 2024.

[3] Matthias Erbar and Jan Maas. Ricci curvature of finite markov chains via convexity of the entropy. *Archive for Rational Mechanics and Analysis*, 206(3):997–1038, August 2012.

[4] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(24):695–709, 2005.

[5] Jan Maas. Gradient flows of the entropy for finite markov chains, 2011.

[6] Jascha Sohl-Dickstein, Peter Battaglino, and Michael Robert DeWeese. A new method for parameter estimation in probabilistic models: Minimum probability flow. *CoRR*, abs/2007.09240, 2020.