Score Matching as Wasserstein Gradient Minimization

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

June 2025

1 Introduction

Score matching, introduced in [3], 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.

This framework can be generalized to different functionals and may lead to new methods for parametric statistical inference.

Finally we will show a deep connection between score matching and minimum probability flow, introduced in [4], that is a different method for estimating statistical models, initially developed for discrete domains such as Ising models.

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 $X^{(1)}, ..., X^{(n)} \overset{\text{iid}}{\sim} \mu_g$, where μ_g is the real (unknown) distribution whose density with respect to the Lebesgue measure is g.
We want to model the distribution with π_θ, absolutely continuous with respect to Lebesgue, and with density $f_\theta(x) = \frac{1}{Z} e^{-H_\theta(x)}$.

Score matching reads

$$\begin{aligned}\theta_{SM} &= \arg\min_\theta \mathbb{E}_g[\|\nabla_x \log f_\theta - \nabla_x \log g\|^2] \\ &= \arg\min_\theta \mathbb{E}_g[(\nabla_x H_\theta)^2 - 2\Delta_x H_\theta],\end{aligned}$$

whose sample version is

$$\begin{aligned}\theta^*_{SM} &= \arg\min_\theta \mathbb{E}_D[\|\nabla_x \log f_\theta - \nabla_x \log g\|^2] \\ &= \arg\min_\theta \mathbb{E}_D[(\nabla_x H_\theta)^2 - 2\Delta_x H_\theta],\end{aligned}$$

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

In [3] 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 this manuscript.

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), 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.

Clearly if a functional $\mathcal{F}$ is displacement convex, then it has a unique minima μ^*.

The main result for this section, denoted in [2] as Poljak–Łojasiewicz inequality, states that the Wasserstein gradient flow with respect to a (strictly)

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 (strictly) displacement convex, so it is reasonable trying to minimize it via Wasserstein gradient flows.

4 Score Matching and KL divergence

Let us come back to our usual framework of μ_g and π_θ. The idea is that if $\mu_g \approx \pi_\theta$, then

$$\mathcal{D}_{KL}(\mu_g\|\pi_\theta) \approx 0, \tag{5}$$

so that the Wasserstein gradient flow of $\mathcal{D}_{KL}(\cdot\|\pi_\theta)$ (which is playing the role of $\mathcal{F}[\cdot]$) started from μ_g will be almost stationary. For a sample $X_0 \sim \mu_g$, since $\dot{X}_0 = -\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\mu_g||\pi_\theta)(X_0)$ is the starting point of the *most economical* ODE inducing (3), the requirement (5) naturally translates into finding

$$\theta^* := \arg\min_\theta \mathbb{E}_g[(-\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\mu_g||\pi_\theta))^2]. \tag{6}$$

But we know that $\nabla_{\mathcal{W}_2}\mathcal{D}_{KL}(\cdot\|\pi_\theta) = \nabla_x H_\theta + \nabla_x \log g$ (consult the other manuscript), so that (6) is aiming to find

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

that is precisely the starting point of score matching.

5 Connections with Minimum Probability Flow

Minimum probability flow, introduced in [4], is another method for statistical inference. In the continuous framework, in can be summarized as $\theta_{MPF} = \arg\min_\theta \mathbb{E}_g[|\frac{d}{dt}\mathcal{D}_{KL}(\mu_t\|\mu_g)|_{t=0}|]$, with μ_t following the Fokker-Planck equation with potential π_θ. (NEEDS TO BE SHOWN PROPERLY, IT IS NOT TRIVIAL FROM THE DEFINITION TO USE THIS AS CHARACTERIZATION).

It has been shown in [4] (using long calculations) that the continuous version

of MPF is equivalent to score matching.

Here we want to show it in another way.
The idea is that the Wasserstein gradient flow of $\mathcal{D}_{KL}(\cdot\|\pi_\theta)$ is the Fokker Planck equation with potential H_θ, and we can show that score matching is in turn equivalent to minimizing $\mathbb{E}_g[|\frac{d}{dt}\mathcal{D}_{KL}(\mu_t\|\pi_\theta)|_{t=0}|]$, as μ_t satisfies the Fokker Planck equation with potential H_θ.

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

[4] 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.