Score Matching and Wasserstein Gradient Flows

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 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 $x_1, ..., x_n \overset{\text{iid}}{\sim} \mu_g$, where $x_i \in \mathbb{R}^d$ and 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 $f_\theta(x) = \frac{1}{Z}e^{-V_\theta(x)}$.

Score matching reads

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

whose sample version is

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

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 the appendix [1].

3.2 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

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

Our next idea is to 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

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

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

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 displacement convex $\mathcal{F}$, started at any $\mu_0 \in \mathcal{P}_2^{ac}(\lambda)$, converges exponentially fast towards the unique minimizer $\mu^* \in \mathcal{P}_2^{ac}(\lambda)$.

4 Score Matching and KL divergence

The main idea is to

5 Similarities with Minimum Probability Flow

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.