The Wasserstein Proximal Gradient Algorithm

Adil Salim[1], Anna Korba[2], Giulia Luise[3]

[1]Visual Computing Center, KAUST. [2]Gatsby Unit, University College London. [3] Department of Computer Science, University College London.

Goal

We propose an optimization algorithm, the Wasserstein Proximal Gradient algorithm (WPG), to solve

$$\min_{\mu\in\mathcal{P}_2} \int F d\mu + \mathcal{H}(\mu),$$

where $\mathcal{P}_2$ is the space of probability measures over $\mathsf{X} = \mathbb{R}^d$ with finite second moment, F is a smooth convex function over X, and $\mathcal{H}$ is a geodesically convex functional over $\mathcal{P}_2$.

We prove convergence rates for the WPG.

The WPG generalizes the proximal gradient algorithm from X to $\mathcal{P}_2$. Its convergence rates generalize those of the proximal gradient algorithm.

The proof relies on viewing WPG as a discretization of a Wasserstein gradient flow, and using geometric insights provided by this point of view.

Background

Wasserstein distance.
Let μ, ν probability measures with finite second moments ($\mu, \nu \in \mathcal{P}_2$).

$$W^2(\nu,\mu) := \inf\{\mathbb{E}\|Y - X\|^2, X \sim \mu, Y \sim \nu\}.$$

If $\mu \ll$ Leb, any minimizer (X, Y) is written $(X, Y) = (X, T_\mu^\nu(X))$.
The Wasserstein space = The metric space $(\mathcal{P}_2, W)$.

JKO operator.
Consider a functional $\mathcal{H} : \mathcal{P}_2 \to (-\infty, +\infty]$ convex along generalized geodesics: for every $\mu, \nu, \pi \in \mathcal{P}_2$ such that $\pi \ll$ Leb and $\alpha \in [0, 1]$,

$$\mathcal{H}(\alpha T_\pi^\mu + (1-\alpha)T_\pi^\nu) \le \alpha\mathcal{H}(\mu) + (1-\alpha)\mathcal{H}(\nu).$$

Then,

$$\mathrm{JKO}_{\mathcal{H}}(\mu) := \arg\min_\nu \mathcal{H}(\nu) + \frac{1}{2}W^2(\mu,\nu) \subset \mathcal{P}_2.$$

Potential.
Consider a smooth convex function F. We define the potential energy $\mathcal{E}_F : \mathcal{P}_2 \to \mathbb{R}$ as,

$$\mathcal{E}_F(\mu) := \int F d\mu.$$

Wasserstein Proximal Gradient Algorithm

We consider the problem

$$\min_{\mu\in\mathcal{P}_2} \mathcal{G} := \mathcal{E}_F + \mathcal{H}. \tag{1}$$

Problem (1) covers several tasks encountered in Machine Learning, for instance the task of sampling w.r.t. the distribution $\propto \exp(-F)$ [3].
We propose WPG, a **natural optimization algorithm** for solving (1):

$$\mu_{n+1} \in \mathrm{JKO}_{\gamma\mathcal{H}}((I - \gamma\nabla F)\#\mu_n),$$

where $\gamma > 0$ and $\#$ is the pushforward operation.

The WPG is natural but not as practical as its concurrents because of the JKO step. However, JKO operators are widely used in numerical analysis. Moreover, we believe that efficient implementations are possible for simple regularizers $\mathcal{H}$ used in ML, as it is the case for many proximity operators over $\mathbb{R}^d$[a].

Intuition: The proximal gradient algorithm

Given a nonsmooth convex function G over X, the proximal gradient algorithm is a standard algorithm to minimize $F + G$. It is written

$$x_{n+1} := \mathrm{prox}_{\gamma G}(x_n - \gamma\nabla F(x_n)),$$

where $\mathrm{prox}_G(x) := \arg\min_{y\in\mathsf{X}} G(y) + \frac{1}{2}\|x - y\|^2$ is the proximity operator of G (note the similarity with the JKO operator).

Moreover, the proximal gradient algorithm can be seen as a discretization of the gradient flow $x'(t) = -\partial(F + G)(x(t))$. Indeed, (x_n) satisfies

$$\frac{x_{n+1} - x_n}{\gamma} \in -\nabla F(x_n) - \partial G(x_{n+1}).$$

It is a **Forward Backward Euler discretization.**

Wasserstein gradient flow

The iterations of WPG are similar to those of the proximal gradient algorithm, therefore one can expect that WPG minimizes (1). Moreover, if $\mathcal{H} = \mathcal{E}_G$, then WPG boils down to the proximal gradient algorithm.

Moreover, WPG can be seen as a Forward Backward Euler **discretization** of some continuous time flow.

Wasserstein gradient flow. The Wasserstein gradient flow $(\mu(t))_t$ of $\mathcal{G}$ is the solution to the system of Evolution Variational Inequalities (EVI) [1]: for every π s.t. $\mathcal{G}(\pi) < \infty$,

$$\frac{d}{dt}W^2(\mu(t),\pi) \le -2\left(\mathcal{G}(\mu(t)) - \mathcal{G}(\pi)\right). \tag{2}$$

In the case $F \equiv 0$, the Wasserstein gradient flow can also be obtained as a continuous time limit of WPG [2].

[a]see www.proximity-operator.net

Contact

adil.salim@kaust.edu.sa, anna.korba@ensae.fr, g.luise.16@ucl.ac.uk.

Main Inequality

We analyze WPG as an optimization algorithm. To this end, we use a numerical analysis point of view. Namely, we prove that WPG satisfies a discrete time version of (2).

If γ is small enough, **for every** $\pi \in \mathcal{P}_2$,

$$W^2(\mu_{n+1},\pi) - W^2(\mu_n,\pi) \le -2(\mathcal{G}(\mu_{n+1}) - \mathcal{G}(\pi)). \tag{3}$$

Along with a descent lemma implying that $(\mathcal{G}(\mu_n))$ is nonincreasing, we obtain convergence rates **similar to those of the proximal gradient algorithm.**

Convergence rates

Main results

Assume that F is convex and L-smooth. Assume that $\mathcal{H}$ is convex along generalized geodesics. Let $\gamma < 1/L$ and $\mu_\star$ a minimizer of $\mathcal{G}$.

Then, $\mathcal{G}(\mu_n) - \mathcal{G}(\mu_\star) \le \frac{W^2(\mu_0,\mu_\star)}{2\gamma n}$.

Moreover, if F is λ-strongly convex, $W^2(\mu_n,\mu_\star) \le (1-\gamma\lambda)^n W^2(\mu_0,\mu_\star)$.

References

[1] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*. Springer Science & Business Media, 2008.

[2] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[3] A. Wibisono. Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, page 2093–3027, 2018.