A Mini-Course on Optimization and Dynamics

From Euclidean Gradient Descent to Wasserstein Gradient Flow

Jia-Jie Zhu

Weierstrass Institute for Applied Analysis and Stochastics, Berlin

August 15, 2023

Euclidean gradient descent

Optimization in $\mathbb{R}^d$

$$\min_{x \in \mathbb{R}^d} f(x).$$

We optimize using the *gradient descent* algorithm

$$x_{k+1} = x_k - \tau_k \cdot \nabla f(x_k)$$

using the "variational principle"

$$x_{k+1} \in \operatorname{argmin}_x f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\tau_k} \|x - x_k\|_2^2$$

From GD to Mirror descent

We define the Bregman divergence associated with ϕ as

$$D_\phi(x, y) = \phi(x) - \phi(y) - \nabla\phi(y)^\top (x - y).$$

Mirror descent update (with quadratic term replaced by Bregman)

$$x_{k+1} = \text{argmin}_x\, f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\tau_k} D_\phi(x, x_k)$$

Example of MD: Euclidean norm

Mirror map $\phi(x) = \frac{1}{2}\|x\|_2^2$.

The resulting mirror descent algorithm

$$x_{k+1} = \text{argmin}_x \; f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2\tau_k}\|x - x_k\|_2^2.$$

This is equivalent to gradient descent with stepsize τ_k.

Example of MD: negative entropy

Mirror map $\phi(x) = \sum_{i=1}^{d} x(i) \log x(i)$. The resulting Bregman divergence is the KL-divergence $D_\phi(x, y) = \sum_{i=1}^{d} x(i) \log \frac{x(i)}{y(i)}$.

If we restrict x to a (discrete probability) simplex $1^\top x = 1$, then the MD update

$$x_{k+1} = \text{argmin}_{1^\top x = 1} \; f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{\tau_k} D_\phi(x, x_k)$$

has the closed-form as *exponentiated gradient*

$$x(i)_{k+1} = \frac{x(i)_k e^{-\tau_k g_k}}{\sum_{j=1}^{n} x(j)_k e^{-\tau_k g_k}}, \quad i = 1, 2, \ldots, n$$

Euclidean gradient descent as discretization of ODE gradient flow

$$x^{k+1} \in \operatorname{argmin}_x \langle \nabla f(x^k), x \rangle_{\mathbb{E}^d} + \frac{1}{2\tau}\|x - x^k\|^2$$

is the explicit Euler scheme for the ODE (for simplicity, we take constant time step τ)

$$\dot{x}(t) = -\nabla f(x(t)).$$

The solution $x(t)$ is an ODE gradient flow and the ODE is the gradient flow equation (GFE). In GF terms, the solution $x(t)$ is also called a **curve of maximal slope** (steepest descent).

Gradient flow dynamics: (nonlinear) ODE

$$\dot{x}(t) = -\nabla f(x(t))$$

$\dot{x}(t) \in X$ provides the **rate (or velocity)** (we can see)

$-\nabla f(x(t)) \in X^*$ provides the **(thermodynamic) force** (can't see; shadow price)

The equation should be written in the **force-balance** form

$$\mathbb{I}_R \dot{x}(t) = -\nabla f(x(t)) \in X^*, \quad \mathbb{I}_R : X \to X^* \text{ is the Riesz isomorphism.}$$

If, in the non-Euclidean setting, $X \ncong X^*$, then we have both force space and rate space GFE.

Energy dissipation balance (equality)

Fenchel-Young For convex ψ, (proof is trivial; $\frac{a^2+b^2}{2} \geq ab$)

$$\psi(x) + \psi^*(\xi) \geq \langle x, \xi \rangle, \forall (x, \xi) \in X \times X^*.$$

Furthermore, if ψ is proper, lsc, and convex, (x^*, ξ^*) is optimal.

By **Fenchel(-Young) duality and optimality**

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) =_{X^*} \langle \nabla f(x(t)), \dot{x} \rangle_X = -\|\nabla f(x(t))\|^2 = -(\frac{1}{2}\|\dot{x}\|^2 + \frac{1}{2}\|\nabla f(x)\|^2)$$

Energy does not necessarily decrease along non-solutions, i.e., only inequality

$$\frac{\mathrm{d}}{\mathrm{d}t} f(z(t)) \geq -(\frac{1}{2}\|\dot{z}\|^2 + \frac{1}{2}\|\nabla f(z(t))\|^2).$$

Evolutionary variational inequality $(\mathrm{EVI})_\lambda$: ODE

Suppose the energy functional f is proper, upper semicontinuous, λ-convex for some $\lambda \in \mathbb{R}$, i.e., f can be non-convex, $\forall s \in [0,1], \forall u_0, u_1 \in \mathbb{R}^d$

$$f((1-s)u_0 + su_1) \leq (1-s)f(u_0) + sf(u_1) - \frac{\lambda}{2}s(1-s)\|u_0 - u_1\|^2.$$

and has compact sublevel sets. Then for any initial condition in the $x(0) \in \mathbb{R}^d$, there exists a unique solution at time t, $x(t) \in \mathbb{R}^d$.

Furthermore, the ODE solution $x(t)$ satisfies $(\mathrm{EVI})_\lambda$, for $t, s \in [0, T]$.

$$\frac{1}{2}\|x(t) - \nu\|^2 \leq \frac{1}{2}\mathrm{e}^{-\lambda(t-s)}\|x(s) - \nu\|^2 + M_\lambda(t-s)(f(\nu) - f(x(t))),$$

$$M_\lambda(\tau) = \int_0^\tau e^{-\lambda(\tau - s)}\,\mathrm{d}s, \quad \forall \nu \in \mathrm{dom}(F) \subset \mathbb{R}^d.$$

Using $(\mathrm{EVI})_\lambda$, we can effortlessly extract convergence results. Suppose a minimizer of the energy exists $x^* \in \operatorname{arginf}_{x\in\mathbb{R}^d} f(x)$, we set $\nu = x^*, s = 0$ in $(\mathrm{EVI})_\lambda$

$$\begin{aligned}\|x(t) - x^*\|^2 &\leq \mathrm{e}^{-\lambda t}\|x(0) - x^*\|^2 + 2M_\lambda(t-s)\left(\inf_{x\in\mathbb{R}^d} f(x) - f(x(t))\right)\\ &\leq \mathrm{e}^{-\lambda t}\|x(0) - x^*\|^2\end{aligned}$$

Gradient flow convergence without (strong) convexity: ODE

Impose the *Polyak-Łojasiewicz* inequality, suppose an optimizing x^* exists

$$\|\nabla f(x(t))\|^2 \geq c \cdot (f(x) - f(x^*)).$$

Starting from EDB

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = -\|\nabla f(x(t))\|^2 \leq -c \cdot (f(x) - f(x^*)) \leq 0,$$

implies exponential convergence of the gradient flow

$$f(x(t)) - f(x^*) \leq e^{-c \cdot t} \left(f(x(0)) - f(x^*) \right).$$

We can also recover the log-Sobolev inequality by setting the energy f as the KL-divergence.

Optimization over probability measures

$$\inf_{\mu \in \mathcal{P}} F(\mu)$$

- $\mathcal{P}$: set of probability measures; the probability "simplex"
- We will work with two types of (probability) measures

$$\mathrm{d}\mu(x) = \rho(x)\ \mathrm{d}x, \quad \mu = \sum_{i \in I} \alpha_i \delta_{x_i}\ \alpha \in \Delta$$

- $M^+ \supseteq \mathcal{P}$: non-negative measures; "cone"
- F: objective function; "energy"

Optimization over probability measures

What can't we just do gradient descent?

$$\mu^{k+1} = \mu^k - \tau_k \cdot \nabla F(\mu^k)$$

- $\nabla F(\mu^k)$ is undefined
- μ^{k+1} must be a probability measure; care needs to be taken
- What can we do instead?

A variational approach

Recall the "variational" formulation of gradient descent

$$x^{k+1} \in \operatorname{argmin}_x \langle \nabla f(x^k), x \rangle_{\mathbb{R}^d} + \frac{1}{2\tau} \|x - x^k\|^2 \iff x_{k+1} = x_k - \tau \cdot \nabla f(x_k)$$

for a suitable τ. This is the variational principle.

Can we do the same for probability measures?

$$\mu^{k+1} \in \operatorname{arginf}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{\tau} \mathcal{D}^2(\mu, \mu^k)$$

for some "distance" measure $\mathcal{D}$. This is sometimes called the *Minimizing Movement Scheme* (MMS).

Variational approach and MMS

$$\mu^{k+1} \in \text{arginf}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{\tau}\mathcal{D}(\mu, \mu^k)$$

We must specify the important ingredients

$$\textit{Energy} : F$$

$$\textit{Geometry} : \mathcal{D}$$

The merit of the right gradient flow formulation of a dissipative evolution equation is that it separates energetics and kinetics: The energetics endow the state space with a functional, the kinetics endow the state space with a (Riemannian) geometry via the metric tensor. [Otto 2001]

Geometry: Wasserstein distance

Definition. The p-Wasserstein distance** between probability measures P, Q on $\mathbb{R}^d$ (with p-th finite moments, $p \geq 1$) is defined through the following Kantorovich problem

$$W_p^p(P, Q) := \inf \left\{ \int |x - y|^p d\Pi(x, y) \,\middle|\, \pi^{(1)}_{\#}\Pi = P,\ \pi^{(2)}_{\#}\Pi = Q \right\}$$

Dual Kantorovich problem

$$W_p^p(P, Q) = \sup \left\{ \int \psi_1(x)\, \mathrm{d}P(x) + \int \psi_2(y)\, \mathrm{d}Q(y) \,\middle|\, \psi_1(x) + \psi_2(y) \leq |x - y|^p \right\}$$

Dynamic formulation: **Benamou–Brenier**

$$W_2^2(P, Q) = \inf \left\{ \int_0^1 \int |v_t|^2 \mathrm{d}\mu_t \mathrm{d}t \,\middle|\, \mu_0 = P, \mu_1 = Q, \frac{\mathrm{d}}{\mathrm{d}t}\mu_t + \mathrm{div}(v_t \mu_t) = 0 \right\}$$

Entropy regularization (Sinkhorn divergence)

$$\inf_{\Pi} \int c(x, y) d\Pi(x, y) + \lambda D_\phi(\Pi \| P \otimes Q)$$

Geometry: (Csizsar) ϕ-divergence

Relative entropy is defined as

$$D_\phi(\mu|\nu) = \begin{cases} \int \phi\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\ \mathrm{d}\nu & \text{if } \mu \ll \nu \\ +\infty & \text{otherwise} \end{cases}$$

We can choose the ϕ functions from the following table to obtain: identity (trivial), Kullback, Hellinger, χ^2

Table 1: Entropy functions, their corresponding reverse entropy, and convex conjugates

Entropy f	f^*	Reverse entropy r	r^*
$f_{\mathrm{Id}}(t) = \begin{cases} 0 & \text{if } t = 1 \\ +\infty & \text{otherwise} \end{cases}$	$f^*_{\mathrm{Id}} = \mathrm{Id}$	$r_{\mathrm{Id}}(t) = \begin{cases} 0 & \text{if } t = 1 \\ +\infty & \text{otherwise} \end{cases}$	$r^*_{\mathrm{Id}} = \mathrm{Id}$
$f_{\mathrm{KL}}(t) = t\log t - t + 1$	$f^*_{\mathrm{KL}}(s) = e^s - 1$	$r_{\mathrm{KL}}(t) = t - 1 - \log t$	$r^*_{\mathrm{KL}}(s) = -\log(1-s)$
$f_{\mathrm{H}}(t) = (\sqrt{t} - 1)^2$	$f^*_{\mathrm{H}}(s) = s/(1-s)$	$r_{\mathrm{H}}(t) = (\sqrt{t} - 1)^2$	$r^*_{\mathrm{H}}(s) = s/(1-s)$
$f_{\chi^2}(t) = (t-1)^2$	$f^*_{\chi^2}(s) = s^2/4 + s$	$r_{\chi^2}(t) = (t-1)^2/t$	$r^*_{\chi^2}(s) = 2 - \sqrt{1-s}$

(table: J. Zhu)

Preliminary: first variation over measures and subdifferentials

The **first variation of a functional** F **at** $\mu \in \mathcal{P}$ is defined as a **function** $\frac{\delta F}{\delta \mu}[\mu]$

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\mu + \epsilon \cdot v)|_{\epsilon=0} = \int \frac{\delta F}{\delta \mu}[\mu](x)\, \mathrm{d}v(x)$$

for any perturbation in measure v such that $\mu + \epsilon \cdot v \in \mathcal{P}$.

The **variational principle**

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} F(\mu + \epsilon \cdot v)|_{\epsilon=0} = 0$$

for all variation v, states the "optimality condition".

We also summon the Fréchet differential on a Banach space X as a set in the dual space

$$DF := \{\xi \in X^* \mid F(\mu) \geq F(\nu) + \langle \xi, \mu - \nu \rangle_X + o\left(\|\mu - \nu\|_X\right) \text{ for } \mu \to \nu\}$$

Three types of energy functionals

Suppose $\mu(x) = \rho(x)\ \mathrm{d}x$ WLOG,

$$\mathcal{F}(\varrho) = \int f(\varrho(x))\mathrm{d}x, \quad \mathcal{V}(\varrho) = \int V(x)\mathrm{d}\varrho, \quad \mathcal{W}(\varrho) = \frac{1}{2}\iint W(x-y)\mathrm{d}\varrho(x)\mathrm{d}\varrho(y)$$

We calculate the first variations (by following the definition)

$$\frac{\delta \mathcal{F}}{\delta \varrho}(\varrho) = f'(\varrho), \quad \frac{\delta \mathcal{V}}{\delta \varrho}(\varrho) = V, \quad \frac{\delta \mathcal{W}}{\delta \varrho}(\varrho) = W * \varrho$$

Back to (discrete-time) gradient flow

Optimization

$$\inf_{\mu \in \mathcal{P}} F(\mu)$$

Recall MMS

$$\mu^{k+1} \in \operatorname{arginf}_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{\tau} \mathcal{D}(\mu, \mu^k)$$

We have specified the important ingredients

$$\textit{Energy} : F, \quad \textit{Geometry} : \mathcal{D}$$

We can construct a concrete instance of MMS for gradient flow by “mix-and-match”.

Wasserstein-MMS: Jordan-Kinderlehrer-Otto (JKO) scheme

a.k.a. Minimizing Movement Scheme (MMS):

$$\mu^{k+1} \in \inf_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu^k)$$

This formulation is very general in the sense that it includes **nonlinear-in-measure** F.
We should think of this as the *gradient descent algorithm for prob. measures*.

Otto's Gradient flow equation in the Wasserstein space

$$\mu^{k+1} \in \inf_{\mu \in \mathcal{P}} F(\mu) + \frac{1}{2\tau} W_2^2(\mu, \mu^k)$$

Continuous-time limit $\tau \to 0$, we have (non-trivially) the **gradient flow equation** (GFE)

$$\partial_t \mu - \nabla \cdot (\mu \nabla \frac{\delta F}{\delta \mu}[\mu]) = 0$$

which describes the dissipation of energy F in $(\mathrm{Prob}(\bar{X}), W_2)$. [Otto et al 90s-2000s, Ambrosio 2005]

In a different flavor, we can write it just like ODE $\dot{x} = -\nabla f(x)$ (in the **rate** form; primal vs. dual force-balance)

$$\partial_t \mu = -\mathbb{K}_{\text{Otto}}(\mu)\ \mathrm{D}F = \nabla \cdot (\mu \nabla \mathrm{D}F).$$

Example: WGF of (Boltzmann/KL/relative) Entropy

nonlinear (in measure) energy (e.g., in variational inference)

$$F(\mu) = \mathrm{D}_{\mathrm{KL}}(\mu \| \pi) = \int \log(\frac{\delta \mu}{\delta \pi}(x))\rho(x)\ \mathrm{d}x$$

$$\frac{\delta F}{\delta \mu}[\mu] = \log \rho - \log \pi,$$

density $\rho := \frac{\mathrm{d}\mu}{\mathrm{d}\mathcal{L}}$ The Fokker-Planck equation as the **Wasserstein gradient flow** [Otto et al. 90s-2000s]

$$\begin{aligned} \partial_t \mu &= \nabla \cdot \left(\mu \nabla \frac{\delta F}{\delta \mu}[\mu] \right) \\ &= \nabla \cdot (\mu(\nabla \log \rho - \nabla \log \pi)) \\ &= \Delta \rho + \nabla \cdot (\rho \nabla \log \pi) \end{aligned}$$

- If π is the Lebesgue measure, we obtain the heat equation $\partial_t \mu = \Delta \mu$
- Note: The force field $\frac{\delta F}{\delta \mu}[\mu]$ and the "score" $\nabla \frac{\delta F}{\delta \mu}[\mu]$ are **not accessible** if μ is atomic. $\implies$ "score-matching"...

Application: sampling and variational inference

Suppose $\pi \propto e^{-V(x)}$, but with unknown normalizing constant, we want

$$\inf_{\mu \in \mathcal{P}} D_{KL}(\mu \| \pi).$$

Using the WGF, we have the Fokker-Planck equation

$$\partial_t \mu = \nabla \cdot (\mu(\nabla \log \rho(x) - \nabla V(x)))$$

Suppose there is a single atom whose state is X_t (R.V.), it is pushed towards the velocity field

$$\nabla \log \rho(X_t) - \nabla V(X_t)$$

We can construct gradient descent

$$X_{t+1} = X_t + \tau \cdot (\nabla \log \rho(X_t) - \nabla V(X_t))$$

Langevin Monte-Carlo forward-Euler discretization

$$X_{t+1} = X_t - \tau \cdot \nabla V(X_t) + \sqrt{2\tau} Z, Z \sim N(0, \mathrm{Id})$$

Application: (distributionally) robust learning with Otto's WGF

We can use our WGF theory (invented 20yr ago; nothing new) to solve Wasserstein DRO for robust learning (also adversarial robustness in [Sinha et al. 2017])

$$\min_{\theta} \sup_{\mu} \mathbb{E}_{\mu}\, l(\theta, x) - \gamma \cdot W_2^2(\mu, \hat{\mu}_N)$$

The inner measure-update step is gradient ascent

$$X_{t+1} = X_t + \tau \nabla l(\theta_t, X_t)$$

where $\tau = \frac{1}{2\gamma}$.Then the whole Wasserstein robust learning is simply gradient descent-ascent (GDA).

Energy dissipation balance of WGF

Recall the ODE case

$$\frac{\mathrm{d}}{\mathrm{d}t} f(x(t)) = -(\frac{1}{2}\|\dot{x}\|^2 + \frac{1}{2}\|\nabla f(x)\|^2)$$

In $(\mathrm{Prob}(\bar{X}), F, W_2)$, **Fenchel(-Young)** yields the **Energy dissipation balance** (equality) [Ambrosio et al. 2007]

$$\frac{\mathrm{d}}{\mathrm{d}t} F(\mu(t)) = -\frac{1}{2}|\mu'|_{W_2}(t)^2 - \frac{1}{2}|\nabla^- F|_{W_2}(\mu(t))^2$$

$$F(\mu(t)) - F(\mu(s)) = -\frac{1}{2}\int_s^t |\mu'|_{W_2}(r)^2 + |\nabla^- F|_{W_2}(\mu(r))^2 \, \mathrm{d}r$$

- metric speed with velocity v_t : $|\mu'|_{W_2}(t) = \sqrt{\int |v_t|^2 \, \mathrm{d}\mu}$
- metric slope: $|\nabla^- F|_{W_2}(\mu(t)) = \sqrt{\int |\nabla \frac{\delta F}{\delta \mu}[\mu](x)|^2 \, \mathrm{d}\mu}$

The velocity field can be identified as $v_t = -\nabla \frac{\delta F}{\delta \mu}[\mu]$. EDB can then be used as the definition of gradient flows (curves of maximal slopes), even without GFE.

For (Boltzmann) entropy $F(u) = \rho \log \rho$, EDB gives $\frac{\mathrm{d}}{\mathrm{d}t} F(\mu(t)) = -\int |\nabla \log \rho|^2 \rho \, \mathrm{d}x$

Evolutionary variational inequality $(\text{EVI})_\lambda$: Wasserstein GF

Under a few technical assumptions and the so-called λ-geodesic-convexity of the energy F, if along a geodesic curve γ,

$$F(\gamma(s)) \leq (1-s)F(\gamma(0)) + sF(\gamma(1)) - \frac{\lambda}{2}s(1-s)W_2^2(\gamma(0),\gamma(1)), \;\forall s \in [0,1].$$

Then, there exists unique gradient flow solution satisfies $(\text{EVI})_\lambda$, for .

$$\frac{1}{2}W_2^2\left(\mu(t),\nu\right) \leq \frac{1}{2}\mathrm{e}^{-\lambda(t-s)}W_2^2\left(\mu(s),\nu\right) + M_\lambda(t-s)(F(\nu) - F(\mu(t))),$$
$$\forall \nu \in \text{dom}(\mathcal{F}), M_\lambda(\tau) = \int_0^\tau e^{-\lambda(\tau-s)}\,\mathrm{d}s.$$

Set $\nu \in \text{arginf}_\mu F(\mu)$, we have exponential convergence in-time and uniqueness of gradient flow.

Thank you!

There are many other active research topics in GF for ML

- Gradient flow structure with *kernel geometry* [also some of my past / current works]
- Unbalanced transport and its gradient flow
- Applications: causal inference, mean-field NN, Nash equilibrium, offline RL, policy optimization. . .

Reference

1. Mielke, A. An introduction to the analysis of gradients systems. Preprint at http://arxiv.org/abs/2306.05026 (2023).
2. Otto, F. The geometry of dissipative evolution equations: the porous medium equation. (2001).
3. Ambrosio, L. & Savaré, G. Gradient Flows of Probability Measures. in Handbook of Differential Equations: Evolutionary Equations vol. 3 1–136 (Elsevier, 2007).
4. Santambrogio, F. Optimal transport for applied mathematicians. Birkäuser, NY 55, 94 (2015).
5. Peletier, M. A. Variational Modelling: Energies, gradient flows, and large deviations. arXiv:1402.1990 [math-ph] (2014).