Efficient displacement convex optimization with particle gradient descent

Hadi Daneshmand[1,2,3], Jason D. Lee[4], Chi Jin[4]

MIT-LIDS[1], Boston University[2], Foundation of Data Science Institute[3], Princeton University[4]

Optimization with interacting particles

Example. Consider sampling from the unit ball in $\mathbb{R}^2$. An indigenous technique is the minimization of

$$F(\mu) := \frac{1}{2}\min \int k(x-y)d\mu(x)d\mu(y).$$

For kernel $k(\Delta) = -\log(\|\Delta\|) + \|\Delta\|^2$, the minimizer is the characteristic function of a ball in R^2 [Frostman, 1935].

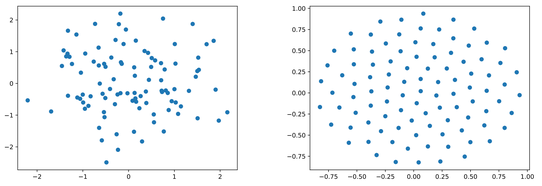
Approximation with particles. The solution to the above variational problem can be approximated by a finite dimensional non-convex program:

$$\min_{w_1,\ldots,w_n} F\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_i}\right),$$

where δ_w is the Dirac measure at w. $w_i \in \mathbb{R}^2$ is called a particle.

Particle gradient descent. One can use gradient descent to optimize the location of particles $w_1, \ldots, w_n \in \mathbb{R}^2$

$$\left\{\forall i : w_i^{(k+1)} = w_i^{(k)} - \gamma\nabla_{w_i}F(\mu_k), \quad \mu_k = \frac{1}{n}\sum_{i=1}^{n}\delta_{w_i^{(k)}}\right\}$$



at initialization after optimization

Figure 1:The distribution of particles before and after optimization with particle gradient descent

Applications.

- Neural networks [Nitanda and Suzuki, 2017]: Each neuron of a two-layer neural network can be viewed as a particle.
- Super-resolution [Chizat and Bach, 2019]: The recovery of sparse functions from its truncated Fourier transform.
- Sampling [Li et al., 2023]: Particles are non-i.i.d. samples that may allow integration over a specific distribution.

Research Goal

Our goal is to establish non-asymptotic approximation guarantees for particle gradient descent when optimizing a specific function class. Consider the sparse measure μ_k with support of size n obtained by k iterations of particle gradient descent. The error for μ_n can be decomposed as

$$F(\mu_k)-F^* := \underbrace{F(\mu_k) - \min_{\mu(n)} F(\mu(n))}_{\text{optimization error}} + \underbrace{\min_{\mu(n)} F(\mu(n)) - F^*}_{\text{approximation error}}.$$

We bound optimization and approximation errors.

Displacement convexity [McCann, 1997]

Non-convexity in particles. Since F is invariant to the permutation of $w_1, \ldots, w_n$, it is not convex in $w_1, \ldots, w_n$. Suppose that $w_1^*, \ldots, w_n^*$ is the unique minimizer of an arbitrary function $F(\frac{1}{n}\sum_{i=1}^n \delta_{w_i})$ such that $w_1^* \neq w_2^*$. If F is invariant to the permutation of $w_1, \ldots, w_n$, then it is non-convex as interpolating $(w_1^*, \ldots w_n^*)$ by $(w_n^*, \ldots w_1^*)$ violates Jensen's inequality.

Definition. Consider two n-sparse distributions μ and ν as

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\delta_{w_i}, \quad \nu = \frac{1}{n}\sum_{i=1}^{n}\delta_{v_i}.$$

Define $\sigma^* = \arg\min_\sigma \sum_{i=1}^n \|w_i - v_{\sigma(i)}\|^2$ where σ is a permutation of $[n]$. Displacement interpolation between μ and ν is defined as

$$\mu_t := \frac{1}{n}\sum_{i=1}^{n}\delta_{tw_i+(1-t)v_{\sigma^*(i)}}$$

F is λ-displacement convex, if

$$F(\mu_t) \leq tF(\mu) + (1-t)F(\nu) - \frac{1}{2}\lambda t(1-t)W_2^2(\mu,\nu)$$

where W_2 is Wasserstein-2 distance (recall $W_2^2(\mu,\nu) := \sum_i \|w_i - v_{\sigma*(i)}\|^2$).

Example 1. *The energy distance between measures over $\mathbb{R}$ is defined as*

$$E(\mu,\nu) = 2\int |x-y|d\mu(x)d\nu(y) - \int |x-y|d\mu(x)d\mu(x) - \int |x-y|d\nu(x)d\nu(y).$$

$E(\mu,\nu)$ is 0-displacement convex in μ [Carrillo et al., 2020].

Weaker than convexity. While convexity asserts Jensen's inequality for $n!$ different possible interpolation of v_i and w_i, displacement convexity only relies on one specific interpolation.

Implications. Consider the following partial differential equation

$$\frac{d\mu}{dt} - div(\mu(t)\partial(dF/d\mu)) = 0$$

If F is λ-displacement convex, then $\mu(t)$ converges to the global minimizer of F at an exponential rate [Santambrogio, 2017].

Optimization error

The optimization error measures how much the function value of μ_n can be reduced by particle gradient descent. Let $\widehat{\mu}$ be a global minimizer of F over n-sparse measures.

Theorem 1. (smooth optimization) *Assume F is ℓ-smooth, and particle gradient descent starts from distinct particles $w_1^{(0)} \neq \ldots \neq w_n^{(0)}$.*

(a) *For $(\lambda > 0)$-displacement functions,*

$$F(\mu_{k+1}) - F(\widehat{\mu}) \leq \ell\left(1 - \left(\frac{2\lambda\ell\gamma}{\ell+\lambda}\right)\right)^k W_2^2(\mu_0,\widehat{\mu})$$

holds as long as $\gamma \leq 2/(\lambda+\ell)$.

(b) *Under 0-displacement convexity,*

$$F(\mu_{k+1}) - F(\widehat{\mu}) \leq \frac{2(F(\mu_0)-F(\widehat{\mu}))W_2^2(\mu_0,\widehat{\mu})}{2W_2^2(\mu_0,\widehat{\mu}) + (F(\mu_0)-F(\widehat{\mu}))\gamma k}$$

holds for $\gamma \leq 1/\ell$.

Optimization error (cont'd)

Convexity vs. displacement convexity. We observe an analogy between the rates for λ-strongly convex functions and λ-displacement convex functions. The main difference is the replacement of Euclidean distance by the Wasserstein distance in the rates for displacement convex functions. This replacement is due to the permutation invariance of F.

Theorem 2. (Lipschitz functions) *Consider the optimization of a L-Lipschitz function with (noisy) particle gradient descent starting from $w_1^{(0)} \neq \ldots \neq w_n^{(0)}$.*

(a) *For λ-displacement functions,*

$$\min_{k\in\{1,\ldots,m\}}\{[F(\mu_k) - F(\widehat{\mu})]\} \leq \frac{2(L^2+1)}{\lambda(m+1)}$$

holds for $\gamma_k = 2/(\lambda(k+1))$.

(b) *If F is 0-displacement convex, then*

$$\min_{k\in\{1,\ldots,m\}}\{[F(\mu_k) - F(\widehat{\mu})]\} \leq \frac{1}{\sqrt{m}}\left(W_2^2(\mu_0,\widehat{\mu}) + L + 1\right)$$

holds for $\gamma_1 = \cdots = \gamma_m = 1/\sqrt{m}$.

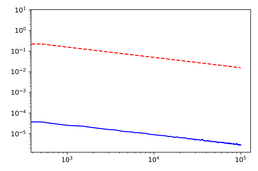


Figure 2:Convergence of noisy particle gradient descent on the energy distance.

Approximation error

The approximation error is induced by the sparsity constraint.

General probabilistic bound. F is L-MMD_K Lipschitz, if there exists a positive definite Kernel K such that

$$|F(\mu) - F(\nu)| \leq L \times \mathrm{MMD}_K(\mu,\nu)$$

holds for all probability. If $\|K\| < B$, then the approximation is bounded by $O(\sqrt{B/n})$ with high probability.

Improved deterministic bound. The approximation error can enjoy a better complexity in n for convex functions.

Lemma 1. *Suppose F is convex and smooth in μ. If the probability measure μ is defined over a compact set, then*

$$\min_{\mu(n)} F(\mu(n)) - F^* = O\left(\frac{1}{n}\right)$$

Remarkably, E in Example 1 is convex in μ.

Applications for neural networks

Consider the class of functions in the following form

$$f(x) = \int \varphi(x^\top w)d\nu(w), \quad \varphi(a) = \begin{cases} 1 & a > 0 \\ 0 & a \leq 0 \end{cases},$$

where $x, w \in \mathbb{R}^2$ lies on the unit circle and ν is a measure with support contained in the upper-half unit circle.

The population loss to optimize approximate μ is defined as

$$\min_{w_1,\ldots,w_n}\left(L(w) := \mathbb{E}_x\left(\frac{1}{n}\sum_{i=1}^{n}\varphi(x^\top w_i) - f(x)\right)^2\right),$$

where x is drawn uniformly from the unit circle. Then, particle gradient descent achievers an $O(n/\sqrt{k} + \frac{1}{n})$ solution with k iterations. Hence, we proves the conjecture of *no optimization-barrier up to permutation invariance*, proposed by [Entezari et al., 2022] for a toy neural network.

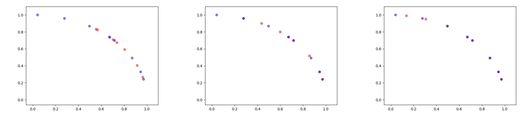


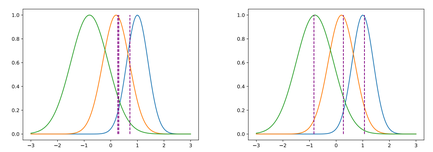
Figure 3:Provable convergence of particle gradients for a toy neural network.

Applications for clustering

Clustering can be formulated as an optimization over sparse measures as

$$\min_{w_1,\ldots,w_n} \mathrm{dist}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_{w_i}, \mu\right)$$

where dist denotes a proper distance metric for probability distributions. Energy distance in Example 2 can be used as a distance [Székely and Rizzo, 2017]. In this case, Theorem 2 provides a non-asymptotic convergence rate to a global optimum.



Initialization After optimization

Figure 4:Clustering in $\mathbb{R}$ with particle gradient descent.

References

[Carrillo et al., 2021] Carrillo, J., Mateu, J., Mora, M., Rondi, L., Scardia, L., and Verdera, J. (2021).
The equilibrium measure for an anisotropic nonlocal energy.
Calculus of Variations and Partial Differential Equations, 60(3):1–28.

[Carrillo et al., 2020] Carrillo, J. A., Francesco, M. D., Esposito, A., Fagioli, S., and Schmidtchen, M. (2020).
Measure solutions to a system of continuity equations driven by newtonian nonlocal interactions.
Discrete and Continuous Dynamical Systems, 40(2):1191–1231.

[Chizat and Bach, 2019] Chizat, L. and Bach, F. (2019).
On the global convergence of gradient descent for over-parameterized models using optimal transport.
Proceedings of Conference on Neural Information Processing Systems.

[Daneshmand and Bach, 2023] Daneshmand, H. and Bach, F. (2023).
Polynomial-time sparse measure recovery: From mean field theory to algorithm design.

[Entezari et al., 2022] Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. (2022).
The role of permutation invariance in linear mode connectivity of neural networks.
ICLR.

[Frostman, 1935] Frostman, O. (1935).
Potentiel d'équilibre et capacité des ensembles.
PhD thesis, Gleerup.

[Li et al., 2023] Li, L., qiang liu, Korba, A., Yurochkin, M., and Solomon, J. (2023).
Sampling with mollified interaction energy descent.
In *The Eleventh International Conference on Learning Representations*.

[McCann, 1997] McCann, R. J. (1997).
A convexity principle for interacting gases.
Advances in mathematics.

[Nitanda and Suzuki, 2017] Nitanda, A. and Suzuki, T. (2017).
Stochastic particle gradient descent for infinite ensembles.
arXiv preprint arXiv:1712.05438.

[Santambrogio, 2017] Santambrogio, F. (2017).
{Euclidean, metric, and Wasserstein} gradient flows: an overview.
Bulletin of Mathematical Sciences.

[Székely and Rizzo, 2017] Székely, G. J. and Rizzo, M. L. (2017).
The energy of data.
Annual Review of Statistics and Its Application.