Proximal Algorithms in Wasserstein Space

Shuailong Zhu

Abstract

This report investigates a basic type of fully backward optimization method – Wasserstein Proximal Algorithm (also known as JKO scheme), in the space of probability measure. Theoretically, we compare the proximal algorithm in Wasserstein Space with its counterpart in Euclidean space to provide some intuitive understanding and some existing convergence rate analysis. We also provide the proof for "geodesic convexity + quadratic growth ⇒ linear convergence". Experimentally, we verify the convergence result of Wasserstein Proximal Algorithms in some simple cases and explore its performance on a two-layer neural network.

1 Introduction

The task of minimizing a functional over the space of probability distributions is common in machine learning, especially in the field of mean-field dynamics of a two-layer neural network [CB18] and generative modeling [Che+24].

Mean-Field Limit - A Particle Perspective. Assume we have a two-layer NN,

$$f(\theta; x) = \frac{1}{m}\sum \varphi(\theta_i; x) = \int \varphi(\theta; x) d\rho_m(\theta) \tag{1}$$

where $\theta_i = (a_i, w_i)$ with $a_i \in \mathbb{R}, b_i \in \mathbb{R}^{d-1}$, $x \in \mathbb{R}^{d-1}$, $y \in \mathbb{R}$, $\rho_m = \frac{1}{m}\sum \delta_{\theta_i}$ and $\varphi(\theta) = a\phi(b^T x)$. Consider a dataset $(x, y) \sim p(x, y)$ and a convex loss $l(\cdot)$,

$$\begin{aligned} R(\rho_m) =& \mathbb{E}_{(x,y)\sim p} l(f(\theta, x), y) \\ =& \mathbb{E}_{(x,y)\sim p} l(\int \varphi(\theta; x) d\rho_m(\theta), y) \end{aligned} \tag{2}$$

and if we assume $m \to \infty$, we can formulate the optimization of two-layer NN as,

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} R(\rho)$$

which is an optimization problem on space of probability measures.

1.1 The Contributions

In this report, we focus on the comparison between the Forward algorithm and Backward Algorithm in Wasserstein Space. We first verify the convergence result of Wasserstein Proximal Algorithms in Langevin dynamics and focus on its performance on two-layer NN optimization.

1.2 The Notations

Let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of probability measure with the finite second moment. The $\mathcal{W}_2$ distance is defined as the square root of,

$$\mathcal{W}_2(\mu, v)^2 = \min_{\gamma \in \Pi(\mu, v)} \int \|x - y\|^2 d\pi(x, y)$$

where $\Pi(\mu, v) \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ whose marginals are μ and v respectively. For a measurable map $T : \mathbb{R}^d \to \mathbb{R}^d$, let $T_\#$ be its pushforward.

Note that $\mathcal{W}_2$ space is not a flat metric space, but is positively curved. We define that F is convex along geodesics defined as: for every $\rho_0, \rho_1 \in \mathcal{P}_2(\mathbb{R}^d)$, F is λ-convex if

$$F(\rho_t) \leq (1 - t)F(\rho_0) + tF(\rho_1) - \frac{1}{2}(1 - t)t\mathcal{W}_2(\rho_0, \rho_1)$$

where ρ_t is the constant speed shortest curve, and can be identified as $((1-t)Id+tT)_{\#}\rho_0$ when $\rho_0 \ll L(\mathbb{R}^d)$. $\lambda = 0$ refers to normal geodesically convexity, and $\lambda > 0$ is called as strongly geodesically convexity.

We also refer to [AGS05] for the concept of convexity along generalized geodesics. Notice that,

$$\text{If } F : \mathcal{P}_2^a(\mathbb{R}^d) \to \mathbb{R}, \text{Convexity along generalized geodesics} \Rightarrow \text{Convexity along geodesic}$$

We also have similar implication under $F : \mathcal{P}_2^a(\mathbb{R}^d)$. However, the definition for convexity along generalized geodesic is stronger in this case [AGS05].

Let $F : \mathcal{P}_2^a(\mathbb{R}^d) \to \mathbb{R}$ be lower semicontinuous and lower bounded. We say that $\xi \in L_2(\rho; \mathbb{R}^d)$ belongs to the Frechet subdifferential $\partial F(\rho)$ of F at ρ if for any $\rho' \in \mathcal{P}_2^a(\mathbb{R}^d)$,

$$F(\rho') \geq F(\rho) + \int_{\mathbb{R}^d} \langle \xi, T_\rho^{\rho'}(\theta) - \theta \rangle d\rho(\theta) + o(W_2(\rho, \rho'))$$

If $\xi \in \partial F(\rho)$ satisfies the following for any transport map T,

$$F(T_{\#}\rho) \geq F(\rho) + \int_{\mathbb{R}^d} \langle \xi, T(\theta) - \theta \rangle d\rho(\theta) + o(\|T - Id\|_{L_2(\rho;\mathbb{R}^d)})$$

then it is called strong subdifferential. $\nabla \dfrac{\delta F}{\delta \rho}(\rho)$ is strong differential where $\dfrac{\delta F}{\delta \rho}(\rho) : \mathbb{R}^d \to \mathbb{R}$ is first variation.

2 Related Works

2.1 Forward and Backward Algorithms in Wasserstein Space

Assume we have a functional $F : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$, our goal is

$$\min_{\rho \in \mathcal{P}_2(\mathbb{R}^d)} F(\rho)$$

Forward Algorithm in Wasserstein Space.

$$\rho_{k+1} = (I - \eta \nabla \frac{\delta F}{\delta \rho}(\rho_k))_{\#}\rho_k \tag{3}$$

where $\nabla \dfrac{\delta F}{\delta \rho}(\rho_k)$ is the Wasserstein differential. It might look difficult to understand at first, but we can understand it from the particle perspective of two-layer NN. If we have a ρ which is a discrete measure, we need to optimize loss $R(\rho)$, then Eqn. 3 can be understood as the gradient descent (GD) for two-layer NN:

$$\theta_i^{k+1} = \theta_i^k - m\eta \nabla R(\frac{1}{m} \sum \delta_{\theta_i^k})$$

Backward Algorithm in Wasserstein Space. JKO scheme is proposed for the optimization in the space of probability measures [JKO98]

$$\rho_{k+1} = JKO_{\eta,F}(\rho_k) \tag{4}$$

where $JKO_{\eta,F}(\rho) = argmin_\rho \, F(\rho) + \dfrac{1}{2\eta}\mathcal{W}^2(\rho_k, \rho)$. It is also known as Wasserstein Proximal Algorithm.

Wasserstein Gradient Flow (WGF). Intuitively, when $\eta \to 0$, both algorithms will converge to a continuity equation,

$$\partial_t \rho_t = \nabla \cdot \left(\rho_t (\nabla \frac{\delta F}{\delta \rho} + \nabla log \rho_t) \right)$$

under certain conditions.

3 Convergence Rate Analysis

3.1 Convergence Rate in Euclidean Space

Before stepping into the optimization in Wasserstein Space, let's have a brief look at Euclidean space. Similarly, we have GD, implicit Euler scheme(Implicit GD), and gradient flow(GF).

Proposition 3.1. *(Informal) The Implicit GD shares a "similar" convergence rate as GF without requiring other assumptions, while GD requires smoothness to ensure the "approximation ability" of first-order approximation.*

Example 1. For instance, if function f is strongly convex, the GF and Implicit GD have linear convergence rates (see Appendix. A.1 for proof), while GD requires smoothness to demonstrate linear convergence which is proved in lectures of Optimization for ML. GF and Implicit GD share a sublinear convergence rate if f is only convex, while smoothness is required for GD to obtain a sublinear rate [Bec17].

Before closing this section, we want to revisit two condition that are essential for obtaining a linear convergence rate for optimizing functions.

$$f(\theta) - f(\theta^*) \geq \frac{1}{2\mu}\|\nabla f(\theta)\|^2 \tag{PL inequality}$$

$$f(\theta) - f(\theta^*) \geq \frac{\mu}{2}\|\theta - \theta^*\|^2 \tag{Quadratic Growth}$$

Simply speaking, QG + convexity ⇒ PL inequality ⇒ linear convergence.

3.2 Convergence Rate in Wasserstein Space

The Wasserstein space case is similar to Euclidean space. Now we consider,

$$F(\rho) = \int V d\rho + \int W(\theta_1, \theta_2) d\rho(\theta_1) d\rho(\theta_2) + \int h(\rho)\rho d\theta$$

The strong differential $\nabla \frac{\delta F}{\delta \rho}(\rho) : \mathbb{R}^d \to \mathbb{R}^d$

$$\nabla \frac{\delta F}{\delta \rho}(\rho) = \nabla \left(V + log\rho + \int W(\theta_1, \cdot) d\rho(\theta_1) + \int W(\cdot, \theta_2) d\rho(\theta_2) \right)$$

Proposition 3.2. *(Informal) If F is strongly convex along generalized geodesics, then WGF of F and Backward Algorithm of F enjoy linear convergence*[1]*.*

Lemma 3.3. *Assume we follow the update rule by Eqn. 4, we have the following two formulas,*

$$\frac{T_{\rho_k}^{\rho_{k-1}} - Id}{\tau} = \nabla \frac{\delta F}{\delta \rho}(\rho_k) = \nabla \left(V + log\rho_k + \int W(\theta_1, \cdot) d\rho_k(\theta_1) + \int W(\cdot, \theta_2) d\rho_k(\theta_2) \right) \tag{5}$$

$$F(\rho_{k-1}) - F(\rho_k) \leq \frac{1}{2\tau} W_2^2(\rho_{k-1}, \rho_k) \tag{6}$$

Theorem 3.4. *Assume V and W convex, $\int h(\rho) d\theta$ is geodesically convex, F satisfies the following λ-quadratic growth condition,*

$$F(\rho) - F(\rho^*) \geq \frac{\lambda}{2} W_2^2(\rho, \rho^*)$$

then linear convergence of the proximal algorithm is guaranteed.

[1] Detailed proof for discrete-time analysis can be found in [Che+24; YY23]. For continuous-time analysis, we can find proof for "geodesically strong convexity in $\mathcal{P}_2^a(\mathbb{R}^d)$ ⇒ linear convergence " in [FGA21].

Proof. By the geodesical convexity of F,

$$\begin{aligned}F(\rho_k) - F(\rho^*) &\leq -\int \langle \nabla \frac{\delta F}{\delta \rho}(\rho_k), T_{\rho_k}^{\rho^*} - Id \rangle d\rho^k \\ &\leq W_2(\rho_k, \rho^*) \sqrt{\int \|\nabla \frac{\delta F}{\delta \rho}(\rho_k)\|^2 d\rho_k} \\ &\leq \frac{1}{\tau} W_2(\rho_k, \rho^*) W_2(\rho_k, \rho_{k-1})\end{aligned}$$

where we use Eqn. 5 for the second "≤". Combined with the QG condition, we have

$$W_2(\rho_k, \rho^*) \leq \frac{2}{\lambda\tau} W_2(\rho_k, \rho_{k-1})$$

Thus

$$F(\rho_k) - F(\rho^*) \leq \frac{2}{\lambda\tau^2} W_2^2(\rho_k, \rho_{k-1})$$

With Eqn. 6,

$$\begin{aligned}F(\rho_{k-1}) - F(\rho_k) &\geq \frac{2}{\lambda\tau^3}(F(\rho^k) - F(\rho^*)) \\ F(\rho_{k-1}) - F(\rho^*) &\geq (1 + \frac{2}{\lambda\tau^3})(F(\rho^k) - F(\rho^*))\end{aligned}$$

□

Corollary 3.5. *If F is KL divergence, geodesically convex, and satisfies LSI inequality, then linear convergence is guaranteed for the proximal algorithm applied to F.*

Proof. LSI is equivalent to Talagland inequality when V is convex, which directly leads to quadratic growth. □

The takeaway is: Backward Algorithm, as discretization of WGF, has fewer requirements on the property of functional. However, Forward Algorithm is easier to compute especially under the particle method setting, while computing the JKO scheme has no closed form in general. As Backward Algorithm is a better approximation for WGF, we would be more interested in its behavior.

Can this paper (Euclidean) [HLO] help with the analysis for semiconvex functional?

4 Methods and Experiments

4.1 Backward Algorithm for Langevin Dynamics

Assume we have a function $V : \mathbb{R}^d \to \mathbb{R}$ and $\int e^{-V} = 1$ for simplicity, the Langevin dynamics is defined as:

$$d\theta_t = \nabla V(\theta_t)dt + \sqrt{2}dW_t$$

This stochastic differential equation can be viewed as a continuous limit of GD with Gaussian noise. The minimizing goal with respect to its continuity equation is,

$$F(\rho) = KL(\rho|e^{-V}) = \int V d\rho + \int log\rho d\rho$$

where the negative entropy regularization comes from the Gaussian noise. When $e^{-V} = \mathcal{N}(0, I)$, $KL(\rho|e^{-V})$ is geodesically strongly convex but non-smooth because of the $\int log\rho d\rho$. Thus, we utilize Backward Algorithm in these situations.

Experiments. We provide numerical experiments to illustrate the dynamical behavior of the Backward Algorithm, similar to [SKL21], where they simulated the Wasserstein Proximal Gradient method for Gaussian distribution[2].

If the initialization distribution and the target distribution are both Gaussian, then we have a closed form of the update [Wib18] and the update will maintain to be Gaussian. Also, we have closed for W_2 distance for two Gaussians known as Bures-Wasserstein distance. We set the initialization distribution to be $\mathcal{N}(0, 100 * I)$ to have the same mean as target distribution which can simplify the update rule as Example 5 in [Wib18], and set the step size $\eta = 0.1$ and iterations 50.

The linear convergence rate is clearly shown in Fig. 1.

[2] Similar to Proximal Gradient Methods in $\mathbb{R}^d$, it is a Forward-Backward method.

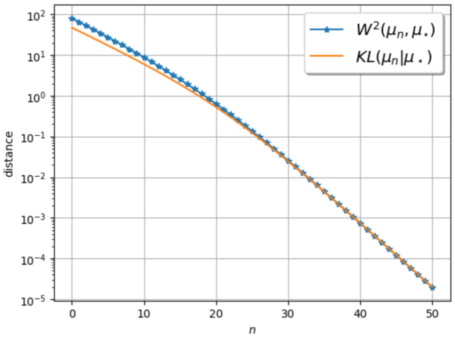


Figure 1: Linear Convergence of Wasserstein distance and KL objective.

4.2 Backward Algorithm for Two-Layer NN

Assume our goal is now:

$$F(\rho) = R(\rho) + \tau H(\rho) \tag{7}$$

where $R()$ is defined as Eqn. 2 and $H()$ is a regularized term. Similar to Langevin dynamics, $H(\rho) = \int log\rho d\rho$ when we apply Gaussian Noisy GD:

$$\theta_i^{k+1} = \theta_i^k - \eta\nabla R(\frac{1}{m}\sum\delta_{\theta_i^k}) + \sqrt{2\eta\tau}z \tag{8}$$

where $z \sim \mathcal{N}(0, I)$. When $\tau \to 0$, Eqn. 8 be equivalent to WGF (the continuity equation) of Eqn. 7. under mild conditions [MMM19]. The main difference compared to previous discussion is, $F(\rho)$ for two-layer NN is convex in the linear combination sense, while previous results like Proposition. 3.2 are based on convexity in the geodesic sense.

Thus, we are interested in: **How would the Backward Algorithm perform in this setting? What is the influence of noise regularization on Forward Algorithms and Backward Algorithms?**

Methods. By the fully backward proximal algorithm, we hope to find T_k such that $T_k\rho_k = \rho_{k+1}$, where ρ_{k+1} minimize the JKO scheme.

$$T^{k+1} = \underset{T}{argmin}\, R(T_\# \rho^k) + \tau H(T_\# \rho^k) + \frac{1}{2\eta}\int \|T(x) - x\|^2 d\rho^k$$

We can estimate the problem using functional approximation,

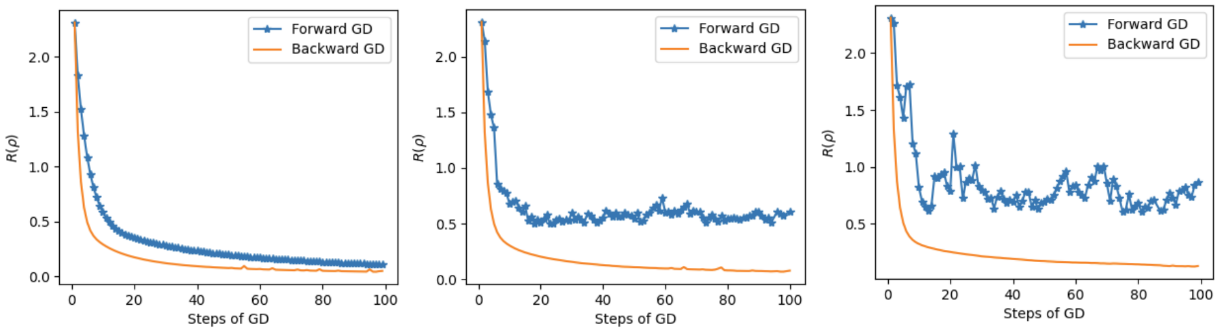
$$\begin{aligned} T^{k+1} = \underset{T}{argmin}\ & \mathbb{E}_{(x,y)\sim p(x,y)}\|\frac{1}{m}\sum_{i=1}^m \varphi(\theta_i^k, x) - y\|^2 \\ & - \frac{\tau}{m}\sum_{i=1}^m log|det\nabla T(\theta_i^k)| + \frac{1}{2m\eta}\sum_{i=1}^m \|T(\theta_i^k) - \theta_i^k\|^2 \end{aligned}$$

where we utilized the change of variable formula [Mok+21]. The idea of functional approximation is to learn one optimal transport map T_k for each iteration.

Experiments. We set the dimension of θ_i to be $d = 3$, number of particles $m = 100$, discretized step size $\eta = 0.1$. We choose activation function $\phi()$ to be $tanh()$ in Eqn. 1, and create a dataset of size 80 by $y = \beta^T x$, where β is fixed and generated with standard Gaussian and x are sampled from standard Gaussian. For the estimation of negative entropy $H(\rho)$, we utilize the nearest neighbor estimator [KL87].

For the training of the T map, we choose a NN of the form $g(\theta) = W_2\sigma(W_1\theta)$, where $W_1 \in \mathbb{R}^{q\times d}$, $W_2 \in \mathbb{R}^{d\times q}$ where $q = 1000$. We use GD optimizer and set the learning rate to be 0.002 and iterations to be 200.

Convergence speed of Backward Algorithm and Forward Algorithm without Gaussian Noise. From Fig. 2 we observe that both Forward and Backward methods can optimize the goal either with or without Gaussian noise. The Backward Algorithm seems to enjoy a faster convergence rate at the beginning of the training, which is

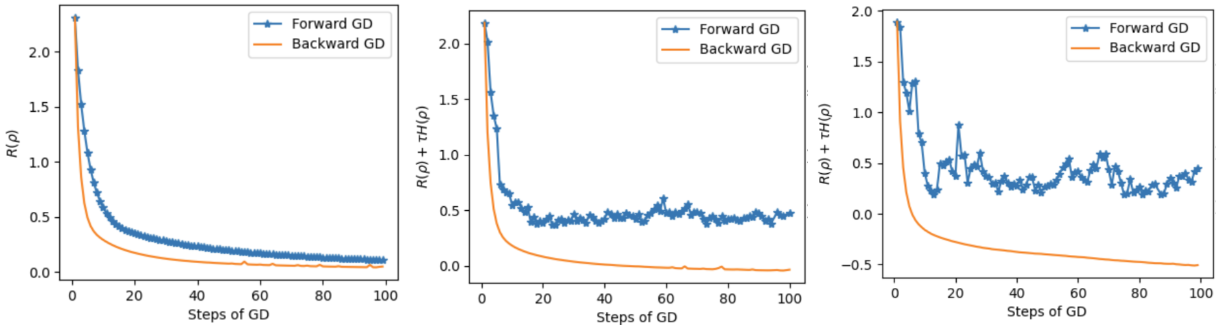


(a) $\tau = 0$

(b) $\tau = 0.03$

(c) $\tau = 0.1$

Figure 2: Evolution of $R(\rho)$ for Two-layer NN with $d = 3$, $m = 100$, and $tanh()$ activation.



(a) $\tau = 0$

(b) $\tau = 0.03$

(c) $\tau = 0.1$

Figure 3: Evolution of $R(\rho) + \tau H(\rho)$ for Two-layer NN with $d = 3$, $m = 100$, and $tanh()$ activation.

a natural benefit from the JKO scheme, as the backward formulation doesn't explicitly constrain the step size and serves a certain local search role.

Convergence Rate Conjecture. The experiments seem to demonstrate a sublinear/linear convergence rate for both algorithms. However, there is no convergence rate result for two-layer NN optimization, which is only convex in linear combination sense. Though there are some global convergence results [CB18; MMM19].

Influence of Noise Regularization. From Fig. 2 and Fig. 3 we can observe that the Backward Algorithm is less sensitive to the Gaussian noise regularization in the aspect of optimizing $R(\rho)$ and $R(\rho) + \tau H(\rho)$.

5 Conclusion

In this project, we applied the Backward Algorithm in Wasserstein Space to a two-layer Neural Network and compared its performance with the Forward Algorithm. Since two-layer NN in the mean-field limit is only convex in the linear combination sense, no existing convergence result exists for it, unlike geodesically convex functional. In experiments, both the Forward Algorithm and Backward Algorithm seem to demonstrate certain sublinear/linear convergence. However, a thorough theoretical analysis is needed.

References

[AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2005.

[Bec17] Amir Beck. *First-Order Methods in Optimization*. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2017. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611974997.

[CB18] Lénaıc Chizat and Francis Bach. "On the Global Convergence of Gradient Descent for Over-parameterized Models using Optimal Transport". In: *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[Che+24] Xiuyuan Cheng et al. *Convergence of flow-based generative models via proximal gradient descent in Wasserstein space*. 2024. arXiv: 2310.17582 [stat.ML].

[FGA21] A. Figalli, F. Glaudo, and European Mathematical Society Publishing House ETH-Zentrum SEW A27. *An Invitation to Optimal Transport, Wasserstein Distances, and Gradient Flows*. EMS textbooks in mathematics. EMS Press, 2021. ISBN: 9783985470105.

[HLO] Tim Hoheisel, Maxime Laborde, and Adam Oberman. "On proximal point-type algorithms for weakly convex functions and their connection to the backward euler method". In: *Optimization Online* ().

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. "The Variational Formulation of the Fokker–Planck Equation". In: *SIAM Journal on Mathematical Analysis* 29.1 (1998), pp. 1–17.

[KL87] Lyudmyla F Kozachenko and Nikolai N Leonenko. "Sample estimate of the entropy of a random vector". In: *Problemy Peredachi Informatsii* 23.2 (1987), pp. 9–16.

[MMM19] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. *Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit*. 2019. arXiv: 1902.06015 [stat.ML].

[Mok+21] Petr Mokrov et al. "Large-Scale Wasserstein Gradient Flows". In: *Advances in Neural Information Processing Systems*. Ed. by A. Beygelzimer et al. 2021.

[SKL21] Adil Salim, Anna Korba, and Giulia Luise. *The Wasserstein Proximal Gradient Algorithm*. 2021. arXiv: 2002.03035 [math.OC].

[Wib18] Andre Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference On Learning Theory*. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 2093–3027.

[YY23] Rentian Yao and Yun Yang. *Mean-field Variational Inference via Wasserstein Gradient Flow*. 2023. arXiv: 2207.08074 [math.ST].

A Convergence Rate in Euclidean Space

A.1 With only Convexity

Now we consider a continuous gradient flow,

$$\frac{d\theta_t}{dt} = -\nabla f(\theta_t)$$

Proposition A.1. *Consider f convex, then GF of f enjoys a sublinear convergence rate.*

Proof.

$$\frac{d}{dt}\frac{1}{2}\|\theta_t - \theta^*\|^2 = -\nabla f(\theta_t)^T(\theta_t - \theta^*) \leq f(\theta^*) - f(\theta_t)$$

$$f(\theta_t) - f(\theta^*) \leq -\frac{d}{dt}\frac{1}{2}\|\theta_t - \theta^*\|^2$$

Take integral of both sides,

$$\frac{1}{T}\int_0^T f(\theta_t)dt - f(\theta^*) \leq \frac{1}{2T}(\|\theta_0 - \theta^*\|^2 - \|\theta_t - \theta^*\|^2)$$

Using Jensen inequality,

$$f(\frac{1}{T}\int_0^T \theta_t dt) - f(\theta^*) \leq \frac{1}{2T}\|\theta_0 - \theta^*\|^2$$

Therefore, we obtain a $O(1/T)$ convergence rate. □

Proposition A.2. *Consider f convex, then the Implicit GD of f enjoys a sublinear convergence rate.*

Proof. Firstly, we have the update rule,

$$\theta_{k+1} = \theta_k - \eta \nabla f(\theta_{k+1})$$

then we have the following two potentially useful formula.

Formula 1.

$$f(\theta^*) - f(\theta_k) \geq \nabla f(\theta_k)^T(\theta^* - \theta_k)$$

Equivalently,

$$f(\theta_k) - f(\theta^*) \leq \nabla f(\theta_k)^T(\theta_k - \theta^*)$$

Formula 2.

$$\begin{aligned} f(\theta_k) - f(\theta_{k+1}) &\geq \nabla f(\theta_{k+1})(\theta_k - \theta_{k+1}) \\ &= \eta \|\nabla f(\theta_{k+1})\|^2 \end{aligned}$$

By the second one, we have,

$$f(\theta_0) - f(\theta_K) \geq \eta \sum_{k=0}^{K-1} \|\nabla f(\theta_{k+1})\|^2$$

Formula 3.

Now we want to explore the third useful one. However the following expansion doesn't work well

$$\begin{aligned} \|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^*\|^2 + \eta^2 \|\nabla f(\theta_{k+1})\|^2 \\ - 2\eta \nabla f(\theta_{k+1})(\theta_k - \theta^*) \end{aligned}$$

Thus we switch to,

$$\begin{aligned} \|\theta_k - \theta^*\|^2 = \|\theta_{k+1} - \theta^*\|^2 + \eta^2 \|\nabla f(\theta_{k+1})\|^2 \\ + 2\eta \nabla f(\theta_{k+1})(\theta_{k+1} - \theta^*) \end{aligned}$$

Thus we have,

$$\begin{aligned} \|\theta_{k+1} - \theta^*\|^2 - \|\theta_k - \theta^*\|^2 &\leq -\eta^2 \|\nabla f(\theta_{k+1})\|^2 - 2\eta(f(\theta_{k+1}) - f(\theta^*)) \\ &\leq 2\eta(f(\theta^*) - f(\theta_{k+1})) \end{aligned}$$

Therefore,

$$f(\theta_{k+1}) - f(\theta^*) \leq \frac{1}{2\eta}\left(\|\theta_k - \theta^*\|^2 - \|\theta_{k+1} - \theta^*\|^2\right)$$

Using Jensen inequality,

$$\begin{aligned} f(\frac{1}{K}\sum_{0}^{K-1} \theta_{k+1}) - f(\theta^*) \leq & \frac{1}{K}\sum [f(\theta_{k+1}) - f(\theta^*)] \\ & \leq \frac{1}{2\eta K}\|\theta_0 - \theta^*\|^2 \end{aligned}$$

We actually only used Formula 1 and Formula 3. □

A.2 With Strong Convexity

Proposition A.3. *Consider f strongly convex and differentiable, then GF of f enjoys linear convergence.*

Proof. By strong convexity,

$$f(\theta) \geq f(\theta_t) + \nabla\langle f(\theta_t), \theta - \theta_t\rangle + \frac{\mu}{2}\|\theta - \theta_t\|^2 \tag{9}$$

Minimize both sides,

$$f(\theta^*) \geq f(\theta_t) - \frac{1}{2\mu}\|\nabla f(\theta_t)\|^2$$

where we let $\theta = \theta_t - \frac{1}{\mu}\nabla f(\theta_t)$ in RHS of Eqn. 9. Thus we will obtain,

$$\frac{df(\theta_t)}{dt} = -\|\nabla f(\theta_t)\|^2 \leq -2\mu\left(f(\theta_t) - f(\theta^*)\right)$$

Therefore,

$$f(\theta_t) - f(\theta^*) \leq e^{-2\mu t}\left(f(\theta_0) - f(\theta^*)\right)$$

□

Proposition A.4. *If f is strongly convex (assume it is also differentiable), then the Implicit GD enjoys linear convergence.*

Proof. Firstly, we have the update rule,

$$\theta_{k+1} = \theta_k - \eta\nabla f(\theta_{k+1})$$

By strong convexity,

$$f(\theta_k) \geq f(\theta_{k+1}) + \langle\nabla f(\theta_{k+1}), \theta_k - \theta_{k+1}\rangle + \frac{\mu}{2}\|\theta_k - \theta_{k+1}\|^2$$

Let $\eta = \frac{C}{\mu}$,

$$\begin{aligned} f(\theta_{k+1}) - f(\theta_k) &\leq -\eta\|\nabla f(\theta_{k+1})\|^2 - \frac{\mu\eta^2}{2}\|\nabla f(\theta_{k+1})\|^2 \\ &= -(\frac{C^2 + 2C}{2\mu})\|\nabla f(\theta_{k+1})\|^2 \end{aligned}$$

Since we have $f(\theta^*) \geq f(\theta_{k+1}) - \frac{1}{2\mu}\|\nabla f(\theta_{k+1})\|^2$,

$$f(\theta_{k+1}) - f(\theta_k) \leq (C^2 + 2C)(f(\theta^*) - f(\theta_{k+1}))$$

Simply, we let $C^2 + 2C = 1$,

$$f(\theta_{k+1}) - f(\theta^*) \leq \frac{1}{2}(f(\theta_k) - f(\theta^*))$$

which implies linear convergence. □

A.3 With only PL Inequality