OPT2019: 11th Annual Workshop on Optimization for Machine Learning

Obtaining Regularization for Free via Iterate Averaging

Jingfeng Wu UUUJF@JHU.EDU
Vladimir Braverman VOVA@CS.JHU.EDU
Johns Hopkins University, Baltimore, MD 21218

Lin F. Yang* LINYANG@EE.UCLA.EDU
University of California, Los Angeles, CA 90095

Abstract

Regularization for optimization is a crucial technique to avoid overfitting in machine learning. In order to obtain the best performance, we usually train a model by tuning the regularization parameters. It becomes costly, however, when a single round of training takes significant amount of time. Very recently, Neu and Rosasco [13] shows that if we run stochastic gradient descent (SGD) on linear regression problems, then by averaging the SGD iterates properly, we obtain a regularized solution. It left open whether the same phenomenon can be achieved for other optimization problems and algorithms. In this paper, we establish a complete theory by showing an averaging scheme that converts the iterates of GD on an arbitrary strongly convex and smooth objective function to its regularized counterpart with an adjustable regularization parameter. Our method can be used for accelerated SGD as well. We derive our results by leveraging the power of approximating the algorithmic path by a continuous differential equation and its discretization. In sum, we obtain regularization *for free* for a large class of optimization problems and resolve an open question in [13].

1. Introduction

Regularization for optimization is a key technique for avoiding over-fitting in machine learning and statistics [6, 11, 18, 19]. The effects of explicit regularization methods, i.e., an extra regularization term added to the vanilla objective, are well studied, e.g., ridge regression [19], LASSO [18] and entropy regularization [6]. Despite the great benefits of adopting explicit regularization, it could cause a huge computational burden to search for the optimal hyper-parameter associated with the extra regularization term, especially for large-scale machine learning problems [5, 8, 14].

In another line of research, people recognize and utilize the implicit regularization caused by certain components in machine learning algorithms, e.g., early stopping [1, 21], different optimization methods [7, 13, 15, 17], iterate averaging [2, 12]. The regularization effect usually happens along the process of running the algorithm and/or requires little post-computation. There is a great deal of evidence indicating that such implicit bias plays a crucial role for the generalization property in many modern machine learning models [10, 22, 24]. However such implicit regularization is often a fixed effect and lacks the ability to be adjusted. To fully utilize such inherent benefits, we need a thorough understanding about the mechanism of the implicit regularization.

* Corresponding author.

Among all the efforts spent on understanding and utilizing the implicit regularization, the work on bridging iterate averaging with explicit regularization [13] is extraordinarily appealing. Concisely, [13] shows that for linear regression, one can achieve the ridge regression (ℓ_2-regularization) effect by simply taking geometrical averaging over the optimization path generated by stochastic gradient descent (SGD), which costs merely a small amount of computation. More interestingly, the regularization is "adjustable", i.e., new regularized solution can be obtained nearly immediately from the stored SGD path, with little additional computation overhead. It avoids the heavy computational demand of hyper-parameter tuning, which is a major disadvantage of the explicit regularization methods.

Nevertheless, [13] only provides a method and its analysis for linear regression optimized by SGD. Since linear regression itself is a rather restricted optimization objective, a nature question arises:

Can we obtain "adjustable" implicit regularization for other objective functions and optimization methods?

In this work, we answer this question positively from two aspects:

1. We show that for Nestrov's accelerated stochastic gradient descent (NSGD), the iterate averaged solution can also realize ℓ_2-regularization by a modified weighting scheme;

2. Beside linear regression, we extend the analysis in [13] and characterize the regularization effects of the iterate averaged solution for strongly convex and smooth loss functions.

Our analysis is motivated from continuous approximation based on differential equations. We discretize the continuous equations and generalize them to algorithms with finite step size. Our results, in addition to the linear regression result in [13], illustrate the promising application of iterate averaging to obtain regularization *for free*, thus providing an efficient method for hyper-parameter tuning with regularization.

2. Preliminaries

Let $\{(x_i, y_i) \in \mathbb{R}^{d\times 1}\}_{i=1}^n$ be the training data. Let $w \in \mathbb{R}^d$ be the parameters to be optimized. For convenience, in the following discussion we always assume w is initialized to zero, i.e., $w_0 = 0$. The loss is denoted as $L(w) = \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, w)$, e.g., for linear regression under the square loss $L(w) = \frac{1}{2n}\sum_{i=1}^n (w^T x_i - y_i)^2$. The explicit regularization term is denoted by $R(w)$, e.g., for ℓ_2-regularization $R(w) = \frac{1}{2}\|w\|_2^2$. Typically, a hype-parameter λ is associated with the regularization term to be balanced with the main loss term, hence the regularized loss becomes $\hat{L}(w) = L(w) + \lambda R(w)$. Given an optimization algorithm, e.g., SGD or NSGD, an optimization path is generated through running the algorithm. With a little abuse of notations, we use w_k to represent the algorithmic iterate at step k of the unregularized loss $L(w)$, while $\hat{w}_k$ for that of the regularized loss $L(\hat{w}) + \lambda R(\hat{w})$, respectively. Sometimes we write $\hat{w}_k$ with a script as $\hat{w}_{k,\lambda}$ to emphasize its dependence on λ.

Stochastic gradient descent In the typical setting of SGD, during every iteration, a mini-batch is randomly sampled, and then parameters update via the gradient of loss estimated by this mini-batch. For simplicity we assume batch-size is 1 and learning rate is constant. Then for the loss $L(w) = \frac{1}{n}\sum_{i=1}^n \ell(x_i, y_i, w)$, with learning rate $\eta > 0$, the SGD takes the following update: $w_{k+1} = w_k - \eta\nabla\ell(x_k, y_k, w_k),\ w_0 = 0$.

Similarly, for the regularized loss $L(\hat{w}) + \lambda R(\hat{w})$, with learning rate $\gamma > 0$, SGD takes update: $\hat{w}_{k+1} = \hat{w}_k - \gamma(\nabla \ell(x_k, y_k, \hat{w}_k) + \lambda \nabla R(\hat{w}_k)),\ \hat{w}_0 = 0$.

Nesterov's accelerated stochastic gradient descent For simplicity we reload the notations used in SGD. Suppose the loss $L(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(x_i, y_i, w)$ is α-strongly convex. Let η be the learning rate and $\tau = \frac{1-\sqrt{\eta\alpha}}{1+\sqrt{\eta\alpha}}$, then NSGD [16, 20] takes update: $w_{k+1} = v_k - \eta \nabla \ell(x_k, y_k, v_k),\ v_k = w_k + \tau(w_k - w_{k-1}),\ w_0 = w_1 = 0$. Thus for linear regression we have,

$$w_{k+1} = v_k - \eta(x_k x_k^T v_k - x_k y_k), \quad v_k = w_k + \tau(w_k - w_{k-1}), \quad w_0 = w_1 = 0. \tag{1}$$

Now consider the loss with ℓ_2-regularization $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_2^2$, which is then $(\alpha + \lambda)$-strongly convex. Thus let $\hat{\tau} = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1+\sqrt{\gamma(\alpha+\lambda)}}$ and γ be the learning rate, then NSGD for the regularized loss takes update: $\hat{w}_{k+1} = \hat{v}_k - \gamma\left(\nabla \ell(x_k, y_k, \hat{v}_k) + \lambda \hat{v}_k\right),\ \hat{v}_k = \hat{w}_k + \hat{\tau}(\hat{w}_k - \hat{w}_{k-1}),\ \hat{w}_0 = \hat{w}_1 = 0$. Specifically for linear regression with ℓ_2-regularization we obtain

$$\hat{w}_{k+1} = \hat{v}_k - \gamma\left((x_k x_k^T + \lambda)\hat{v}_k - x_k y_k\right), \quad \hat{v}_k = \hat{w}_k + \hat{\tau}(\hat{w}_k - \hat{w}_{k-1}), \quad \hat{w}_0 = \hat{w}_1 = 0. \tag{2}$$

Iterate averaging A weighting scheme p_k is defined by a probability distribution associated to the optimization path, i.e., $p_k \geq 0,\ k \geq 0,\ \sum_{k=0}^{\infty} p_k = 1$. Its accumulation is denoted as $P_k = \sum_{i=0}^{k} p_i$, where $\lim_{k\to\infty} P_k = 1$. Given a weighting scheme p_k (or equivalently P_k), the iterate averaged path is defined as

$$\tilde{w}_k = \frac{1}{P_k}\sum_{i=0}^{k} p_i w_i, \quad k \geq 0. \tag{3}$$

The properties of various kinds of averaging schemes (for the SGD optimization path) have been investigated before. E.g., arithmetic average is shown to bring better convergence [2, 12]; tail-averaging is analyzed in [9]; and [13] discusses geometrically averaging and its regularization bias. This work is inspired by [13].

3. Main results

Let us start with revisiting the existing results on connecting iterate averaging with explicit regularization. For linear regression with square loss, [13] shows that if taking geometric iterate averaging over the optimization path of SGD, one can obtain the solution of linear regression regularized by ℓ_2-regularization. In the following, we will generalize their results to 1) NSGD optimization path and 2) strongly convex and smooth losses. Not limited to SGD and linear regression, our results manifest the broader potential of applying iterate averaging in nearly computation-free hyper-parameter tuning, model selection, and regularizing the model, etc.

Our analysis is motivated from continuous differential equations, which is left in Appendix A, due to the space limitation. In the following we introduce our results in discrete cases.

3.1. The regularization effect of iterate averaging over NSGD path

We firstly elaborate our results on the regularization effect realized by NSGD with iterate averaging. Concisely, we show that for linear regression, there exists a certain kind of geometric averaging schemes, such that iterate averaging over NSGD optimization path could obtain the effect of ℓ_2-regularization.

Theorem 1 (The regularization bias of iterate averaging over NSGD path) *Consider linear regression and ℓ_2-regularization. Suppose the loss function is α-strongly convex and β-smooth ($\beta \geq \alpha > 0$). Let the NSGD optimization paths of the unregularized and regularized losses be defined as in Eq. (1) and Eq. (2) respectively. Let $\lambda,\ \eta,\ \gamma$ be such that $\eta = \frac{\gamma}{1-\lambda\gamma}$, $0 < \eta < \frac{1}{\beta}$. If the iterate averaged solution $\tilde{w}_k$ in Eq. (3) is achieved with respect to the weighting scheme*

$$P_k = 1 - \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad k \geq 0, \tag{4}$$

then we have

1. *For all $k \geq 0$,*
$$\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] = (1 - P_k)\left(\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]\right). \tag{5}$$
2. *Both $\mathbb{E}[w_k]$ and $\mathbb{E}[\hat{w}_k]$ converge. Thus for $C = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \in (0,1)$, there exists a constant K such that for all $k > K$,*
$$\left\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\right\|_2 \leq \mathcal{O}(C^k). \tag{6}$$
Hence the limitation of $\mathbb{E}[\tilde{w}_k]$ exists and $\lim_{k\to\infty}\mathbb{E}[\tilde{w}_k] = \lim_{k\to\infty}\mathbb{E}[\hat{w}_k]$.

The proof is deferred to Appendix C.1.

Theorem 1 answers one of the open questions in [13]: with some modifications on the weighting scheme, iterate averaging over NSGD path realizes ℓ_2-regularization, similar to that over SGD path. We provide not only the convergence of $\tilde{w}_k$ to $\hat{w}_k$ (Eq. (6)), but also the finite step convergence error between them (Eq. (5)).

3.2. The regularization effect of iterate averaging for strongly convex and smooth objectives

Now we turn to explore the regularization bias of iterate averaging for more general loss functions, e.g., strongly convex and smooth ones. Suppose the loss function $L(w) = \frac{1}{n}\sum_{i=1}^{n}\ell(x_i, y_i, w)$ is α-strongly convex and β-smooth, $\beta \geq \alpha > 0$, and lower bounded. For clarity, we focus on gradient descent (GD) in this part, i.e.,

$$w_{k+1} = w_k - \eta\nabla L(w_k), \quad \hat{w}_{k+1,\lambda} = \hat{w}_{k,\lambda} - \gamma(\nabla L(\hat{w}_{k,\lambda}) + \lambda\hat{w}_{k,\lambda}), \quad w_0 = \hat{w}_{0,\lambda} = 0. \tag{7}$$

Let $b = -\nabla L(w_0),\ w_0 = 0$. Without loss of generality, let the unique optimal w_* of $L(w)$ satisfies $w_* > w_0 = 0$, entry-wisely. Then we have (see Lemma 5)

$$\alpha w - b \leq \nabla L(w) \leq \beta w - b, \quad w \in (0, w_*), \tag{8}$$

where the "$\leq$" is defined entry-wisely. Let us denote

$$u_{k+1} - u_k = -\eta(\alpha u_k - a), \quad v_{k+1} - v_k = -\eta(\beta v_k - a), \quad u_0 = v_0 = 0, \tag{9}$$

and $\tilde{u}_k = \frac{1}{P_k}\sum_{i=0}^{k} p_i u_i,\ \tilde{v}_k = \frac{1}{P_k}\sum_{i=0}^{k} p_i v_i,\ k \geq 0$. One can view that u_k and v_k control w_k in Eq. (7) from both upper and lower side respectively. With these conventions, we introduce Theorem 2, which characterizes the regularization effect of iterate averaging over GD paths for general strongly convex and smooth losses.

Theorem 2 (The regularization bias of iterate averaging for strongly convex and smooth loss) *Consider α-strongly convex and β-smooth ($\beta \geq \alpha > 0$) loss functions and ℓ_2-regularization. Let the unregularized and regularized GD optimization paths be defined as in Eq. (7). Let $\lambda_1,\ \lambda_2,\ \eta,\ \gamma$ be such that $\lambda_1 = \frac{1}{\gamma} - \frac{1}{\eta} + \beta - \alpha$, $\lambda_2 = \frac{1}{\gamma} - \frac{1}{\eta} + \alpha - \beta$, $\frac{1}{2\beta-\alpha} < \eta < \frac{1}{\beta}$, $0 < \gamma < \frac{1}{\beta-\alpha+1/\eta}$. If the iterate averaged solution $\tilde{w}_n$ in Eq. (3) is achieved with respect to the weighting scheme $P_k = 1 - \left(\frac{\gamma}{\eta}\right)^{k+1}$, then we have*

1. *For all $k \geq 0$,*

$$\hat{w}_{k,\lambda_1} + (1 - P_k)(\tilde{v}_k - v_k) \leq \tilde{w}_k \leq \hat{w}_{k,\lambda_2} + (1 - P_k)(\tilde{u}_k - u_k), \tag{10}$$

   where the "$\leq$" is defined entry-wisely.

2. *$u_k, \tilde{u}_k, v_k, \tilde{v}_k, \hat{w}_{k,\lambda_1}, \hat{w}_{k,\lambda_2}$ converge. Thus let $m = (\hat{w}_{\infty,\lambda_2} + \hat{w}_{\infty,\lambda_1})/2$, $d = (\hat{w}_{\infty,\lambda_2} - \hat{w}_{\infty,\lambda_1})/2$, there exist constants $C = \max\{(1 - \gamma(\alpha + \lambda_1), (1 - \gamma(\alpha + \lambda_2), \frac{\gamma}{\eta}\} \in (0, 1)$ and K, such that for all $k > K$ we have*

$$\|\tilde{w}_k - m\|_2 \leq \|d\|_2 + \mathcal{O}(C^k). \tag{11}$$

The proof is deferred to Appendix C.2.

According to Theorem 2, for strongly convex and smooth functions, the geometrically averaged GD path $\tilde{w}_k$ lies around the area between two regularized GD paths, $\hat{w}_{k,\lambda_1}$ and $\hat{w}_{k,\lambda_2}$. Furthermore, it converges to a cube with diagonal vertices as $\hat{w}_{\infty,\lambda_1}$ and $\hat{w}_{\infty,\lambda_2}$. In this way we predict the performance of the regularized solution with hyper-parameter between λ_2 and λ_1. This result can potentially benefit the process of searching the hyper-parameter associated with ℓ_2-regularization.

4. Conclusions

In this work we show that there exists a certain type of weighting scheme such that the iterate averaged solution can realize ℓ_2-regularization, for Nesterov's accelerated stochastic gradient descent with quadratic loss functions, and gradient descent with strongly convex and smooth objectives. This in particular resolves an open question in [13]. Our results demonstrate the potential of adopting iterate averaging to obtain regularization for free in a much broader class of optimization objectives and optimization methods.

References

[1] Alnur Ali, J Zico Kolter, and Ryan J Tibshirani. A continuous-time view of early stopping for least squares regression. *arXiv preprint arXiv:1810.10082*, 2018.

[2] Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate o (1/n). In *Advances in neural information processing systems*, pages 773–781, 2013.

[3] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[4] Dean S Clark. Short proof of a discrete gronwall inequality. *Discrete applied mathematics*, 16 (3):279–281, 1987.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[6] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2005.

[7] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. *arXiv preprint arXiv:1802.08246*, 2018.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2016. doi: 10.1109/cvpr.2016.90. URL http://dx.doi.org/10.1109/CVPR.2016.90.

[9] Prateek Jain, Sham Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: mini-batching, averaging, and model misspecification. *Journal of Machine Learning Research*, 18, 2018.

[10] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

[11] Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

[12] Chandrashekar Lakshminarayanan and Csaba Szepesvari. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355, 2018.

[13] Gergely Neu and Lorenzo Rosasco. Iterate averaging as regularization for stochastic gradient descent. *arXiv preprint arXiv:1802.08009*, 2018.

[14] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354, 2017.

[15] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research*, 19(1):2822–2878, 2018.

[16] Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

[17] Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. In *Advances in Neural Information Processing Systems*, pages 10608–10619, 2018.

[18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[19] Andrey N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of ill-posed problems*. V. H. Winston & Sons, Washington, D.C.: John Wiley & Sons, New York, 1977. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics.

[20] Lin Yang, Raman Arora, Tuo Zhao, et al. The physical systems behind optimization algorithms. In *Advances in Neural Information Processing Systems*, pages 4372–4381, 2018.

[21] Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[22] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

[23] Xingyu Zhou. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

[24] Zhanxing Zhu, Jingfeng Wu, Bing Yu, Lei Wu, and Jinwen Ma. The anisotropic noise in stochastic gradient descent: Its behavior of escaping from minima and regularization effects. *arXiv preprint arXiv:1803.00195*, 2018.

Appendix A. Continuous analysis

A.1. Preliminaries

To motivate our proofs for the theorems in main text, let us first elaborate the continuous cases, then we will extend our analysis to the discrete circumstances. One may ignore this part and go directly to Appendix C for the missing proofs in main text, which is self-consistent.

Continuous optimization paths To ease notations and preliminaries, in this section we only discuss gradient descent (GD) and Nesterov's accelerated gradient descent (NGD), and their strong continuous approximation via ordinary differential equations (ODEs).

We consider loss $L(w)$ and ℓ_2-regularization, $R(w) = \frac{1}{2}\|w\|_2^2$. Let the learning rate $\eta \to 0$, the path of $L(w)$ optimized by GD converges to the following ODE [20]

$$\mathrm{d}w_t = -\nabla L(w_t)\,\mathrm{d}t. \tag{12}$$

Similarly the continuous GD optimization path of regularized loss is

$$\mathrm{d}\hat{w}_t = -[\nabla L(\hat{w}_t) + \lambda \hat{w}_t]\,\mathrm{d}t. \tag{13}$$

As for NGD, [16, 20] show if the loss is α-strongly convex, then the NGD optimization path converges to

$$w_t'' + 2\sqrt{\alpha} w_t' + L'(w_t) = 0. \tag{14}$$

Since $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_2^2$ is $(\alpha+\lambda)$-strongly convex, the NGD path of the regularized loss satisfies

$$\hat{w}_t'' + 2\sqrt{\alpha+\lambda}\hat{w}_t' + L'(\hat{w}_t) + \lambda \hat{w}_t = 0. \tag{15}$$

Continuous weighting scheme We define the continuous weighting scheme as

$$p_t \geq 0, \quad t \geq 0, \quad P_t = \int_0^t p(s)\,\mathrm{d}s, \quad \lim_{t\to\infty} P_t = 1. \tag{16}$$

Lemma 3 *Given two continuous dynamic x_t, $\hat{x}_t$, $t \geq 0$. Let $\tilde{x}_t = \frac{1}{P_t}\int_0^t p_s x_s\,\mathrm{d}s$. Suppose $x_0 = \hat{x}_0 = 0$. If the continuous weighting scheme P_t satisfies*

$$\mathrm{d}\hat{x}_t = (1 - P_t)\,\mathrm{d}x_t, \quad t \geq 0, \tag{17}$$

then we have

$$P_t(x_t - \tilde{x}_t) = x_t - \hat{x}_t, \quad t \geq 0, \tag{18}$$

and

$$\hat{x}_t - \tilde{x}_t = (1 - P_t)(x_t - \tilde{x}_t), \quad t \geq 0. \tag{19}$$

Proof By definition we have for $t \geq 0$,

$$\begin{aligned}
\tilde{x}_t &= \frac{1}{P_t}\int_0^t p_s x_s\,\mathrm{d}s = \frac{1}{P_t}\left[x_s P_s|_0^t - \int_0^t P_s\,\mathrm{d}x_s\right] = x_t - \frac{1}{P_t}\int_0^t P_s\,\mathrm{d}x_s \\
&= x_t - \frac{1}{P_t}\left[x_t - \int_0^t (1 - P_s)\,\mathrm{d}x_s\right] = x_t - \frac{1}{P_t}\left[x_t - \int_0^t \mathrm{d}\hat{x}_s\right] \\
&= x_t - \frac{1}{P_t}(x_t - \hat{x}_t).
\end{aligned} \tag{20}$$

Thus

$$P_t(x_t - \tilde{x}_t) = x_t - \hat{x}_t, \tag{21}$$

and

$$\hat{x}_t - \tilde{x}_t = x_t - P_t(x_t - \tilde{x}_t) - \tilde{x}_t = (1 - P_t)(x_t - \tilde{x}_t). \tag{22}$$

These conclude our proof. ■

A.2. Continuous Theorem 1

In this part we study linear regression $L(w) = \frac{1}{2n}\sum_{i=1}^n (w^T x_i - y_i)^2 = \frac{1}{2} w^T \Sigma w - w^T a + \text{const}$ where $\Sigma = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$, $a = \frac{1}{n}\sum_{i=1}^n x_i y_i$, and ℓ_2-regularization $R(w) = \frac{1}{2}\|w\|_2^2$ and continuous NGD paths. Assume the initial condition $w_0 = w_0' = 0$ and $\hat{w}_0 = \hat{w}_0' = 0$. According to Eq. (14) and Eq. (15) the unregularized and regularized NGD dynamics are

$$w_t'' + 2\sqrt{\alpha} w_t' + \Sigma w_t - a = 0, \quad w_0 = w_0' = 0, \tag{23}$$

and

$$\hat{w}_t'' + 2\sqrt{\alpha + \lambda}\hat{w}_t' + (\Sigma + \lambda)\hat{w}_t - a = 0, \quad \hat{w}_0 = \hat{w}_0' = 0. \tag{24}$$

We first solve the order-2 ODE Eq. (23) in the canonical way, and then obtain the solution of Eq. (24) similarly. To do so, let's firstly ignore the constant term and solve the homogenous ODE of Eq. (23), and obtain two general solutions of the homogenous equation as

$$w_{t,1} = e^{\sqrt{\alpha} t} \cos\sqrt{\Sigma - \alpha} t, \quad w_{t,2} = e^{\sqrt{\alpha} t} \sin\sqrt{\Sigma - \alpha} t. \tag{25}$$

Then we guess a particular solution of Eq. (23) as $w_{t,0} = \frac{a}{\Sigma}$. Thus the general solution of ODE (23) can be decomposed as $w_t = \lambda_1 w_{t,1} + \lambda_2 w_{t,2} + w_{t,0}$. Consider the initial conditions $w_0 = w_0' = 0$, we obtain $\lambda_1 = -\frac{a}{\Sigma}$, $\lambda_2 = -\frac{a}{\Sigma}\sqrt{\frac{\alpha}{\Sigma - \alpha}}$. Thus the solution of Eq. (23) is

$$\begin{aligned} w_t &= \frac{a}{\Sigma}\left[1 - e^{-\sqrt{\alpha} t}\cos\sqrt{\Sigma - \alpha} t - \sqrt{\frac{\alpha}{\Sigma - \alpha}} e^{-\sqrt{\alpha} t}\sin\sqrt{\Sigma - \alpha} t\right], \\ w_t' &= \frac{a}{\sqrt{\Sigma - \alpha}} e^{-\sqrt{\alpha} t}\sin\sqrt{\Sigma - \alpha} t. \end{aligned} \tag{26}$$

Repeat these procedures, Eq. (26) is solved by

$$\begin{aligned} \hat{w}_t &= \frac{a}{\Sigma + \lambda}\left[1 - e^{-\sqrt{\alpha+\lambda} t}\cos\sqrt{\Sigma - \alpha} t - \sqrt{\frac{\alpha + \lambda}{\Sigma - \alpha}} e^{-\sqrt{\alpha+\lambda} t}\sin\sqrt{\Sigma - \alpha} t\right], \\ \hat{w}_t' &= \frac{a}{\sqrt{\Sigma - \alpha}} e^{-\sqrt{\alpha+\lambda} t}\sin\sqrt{\Sigma - \alpha} t. \end{aligned} \tag{27}$$

Now let the continuous weighting scheme be

$$P_t = 1 - e^{-(\sqrt{\alpha+\lambda} - \sqrt{\lambda})t}, \tag{28}$$

then we have

$$\mathrm{d}\hat{w}_t = (1 - P_t)\,\mathrm{d}w_t, \tag{29}$$

thus by Lemma 3 we obtain

$$\hat{w}_t - \tilde{w}_t = (1 - P_t)(w_t - \tilde{w}_t), \tag{30}$$

which is the continuous version of the equality in Theorem 1.

A.3. Continuous Theorem 2

Now let's consider α-strongly convex and β-smooth loss function $L(w)$, and ℓ_2-regularization. First adopting the assumption in Theorem 2, i.e., $w_* > w_0 = 0$, and by Lemma 5 we have

$$\alpha w - b \le \nabla L(w) \le \beta w - b, \quad w \in (0, w_*), \tag{31}$$

where $w_0 = 0, b = -\nabla L(0)$, and "$\le$" is defined entry-wisely. We study the continuous optimization paths caused by GD.

Consider the following three dynamics:

$$\mathrm{d}w_t = -\nabla L(w_t)\,\mathrm{d}t, \quad \mathrm{d}u_t = -(\alpha u_t - b)\,\mathrm{d}t, \quad \mathrm{d}v_t = -(\beta v_t - b)\,\mathrm{d}t, \quad w_0 = u_0 = v_0 = 0. \tag{32}$$

By the comparison theorem of ODEs (Gronwall's inequality), and solution of linear ODEs, we claim that for all $t > 0$,

$$v_t \le w_t \le u_t, \quad u_t = \frac{b}{\alpha}(1 - e^{-\alpha t}), \quad v_t = \frac{b}{\beta}(1 - e^{-\beta t}). \tag{33}$$

In a similar manner, for the following three dynamics of regularized loss:

$$\begin{aligned} &\mathrm{d}\hat{w}_{t,\lambda} = -(\nabla L(\hat{w}_{t,\lambda}) + \lambda \hat{w}_{t,\lambda})\,\mathrm{d}t, \\ &\mathrm{d}\hat{u}_{t,\lambda} = -((\lambda+\alpha)\hat{u}_{t,\lambda} - b)\,\mathrm{d}t, \quad \mathrm{d}\hat{v}_{t,\lambda} = -((\lambda+\beta)\hat{v}_{t,\lambda} - b)\,\mathrm{d}t, \end{aligned} \tag{34}$$

where $\hat{w}_{0,\lambda} = \hat{u}_{0,\lambda} = \hat{v}_{0,\lambda} = 0$. Similarly we have for all $t > 0$,

$$\hat{v}_{t,\lambda} \le \hat{w}_{t,\lambda} \le \hat{u}_{t,\lambda}, \quad \hat{u}_{t,\lambda} = \frac{b}{\lambda+\alpha}(1 - e^{-(\lambda+\alpha)t}), \quad \hat{v}_{t,\lambda} = \frac{b}{\lambda+\beta}(1 - e^{-(\lambda+\beta)t}). \tag{35}$$

For the continuous weighting scheme

$$P_t = 1 - e^{-\zeta t}, \quad p_t = \zeta e^{-\zeta t}, \quad t \ge 0, \quad \zeta > 0, \tag{36}$$

the averaged solution is defined as $\tilde{w}_t = \frac{1}{P_t}\int_0^t p_t w_t\,\mathrm{d}t = w_t - \frac{1}{P_t}\int_0^t P_s\,\mathrm{d}w_s$, similar there are $\tilde{u}_t, \tilde{v}_t$. Thanks to Eq. (33) and p_t being non-negative, we have $\tilde{v}_t \le \tilde{w}_t \le \tilde{u}_t$. Let

$$\lambda_1 = \zeta + \beta - \alpha, \quad \lambda_2 = \zeta + \alpha - \beta, \tag{37}$$

then

$$\begin{aligned} P_t(u_t - \tilde{u}_t) = \int_0^t P_s\,\mathrm{d}u_s &= \int_0^t (1 - e^{-(\lambda_2+\beta-\alpha)s}) b e^{-\alpha s}\,\mathrm{d}t = b\int_0^t e^{-\alpha s} - e^{-(\beta+\lambda_2)s}\,\mathrm{d}s \\ &= b\left(\frac{1}{\alpha}(1 - e^{-\alpha t}) - \frac{1}{\lambda_2+\beta}(1 - e^{-(\lambda_2+\beta)t})\right) = u_t - \hat{v}_{t,\lambda_2}. \end{aligned} \tag{38}$$

Thus

$$\tilde{w}_t - \hat{w}_{t,\lambda_2} \le \tilde{u}_t - \hat{v}_{t,\lambda_2} = \tilde{u}_t - u_t + P_t(u_t - \tilde{u}_t) = (1 - P_t)(\tilde{u}_t - u_t). \tag{39}$$

Similarly, since

$$\begin{aligned} P_t(v_t - \tilde{v}_t) = \int_0^t P_s\,\mathrm{d}v_s &= \int_0^t (1 - e^{-(\lambda_1-\beta+\alpha)s}) b e^{-\beta s}\,\mathrm{d}t = b\int_0^t e^{-\beta s} - e^{-(\alpha+\lambda_1)s}\,\mathrm{d}s \\ &= b\left(\frac{1}{\beta}(1 - e^{-\beta t}) - \frac{1}{\lambda_1+\alpha}(1 - e^{-(\lambda_1+\alpha)t})\right) = v_t - \hat{u}_{t,\lambda_1}, \end{aligned} \tag{40}$$

we can obtain a lower bound as

$$\tilde{w}_t - \hat{w}_{t,\lambda_1} \geq \tilde{v}_t - \hat{u}_{t,\lambda_1} = \tilde{v}_t - v_t + P_t(v_t - \tilde{v}_t) = (1 - P_t)(\tilde{v}_t - v_t). \tag{41}$$

These inequalities give us

$$\hat{w}_{t,\lambda_1} + (1 - P_t)(\tilde{v}_t - v_t) \leq \tilde{w}_t \leq \hat{w}_{t,\lambda_2} + (1 - P_t)(\tilde{u}_t - u_t), \tag{42}$$

which is the continuous version of the inequality in Theorem 2.

Appendix B. Technical Lemmas

Lemma 4 *Given two series $x_k,\ \hat{x}_k,\ k \geq 0$, let $\tilde{x}_k = \frac{1}{P_k}\sum_{i=0}^{k} p_i x_i$. Suppose $x_0 = \hat{x}_0 = 0$. If the weighting scheme P_k satisfies*

$$\hat{x}_{k+1} - \hat{x}_k = (1 - P_k)(x_{k+1} - x_k), \quad k \geq 0, \tag{43}$$

then we have

$$P_k(x_k - \tilde{x}_k) = x_k - \hat{x}_k, \quad k \geq 0, \tag{44}$$

and

$$\hat{x}_k - \tilde{x}_k = (1 - P_k)\,(x_k - \tilde{x}_k), \quad k \geq 0. \tag{45}$$

Proof By definition we know $p_0 = P_0,\ p_k = P_k - P_{k-1},\ k \geq 1$, and

$$\begin{aligned} P_k\tilde{x}_k = \sum_{i=1}^{k} p_i x_i = \sum_{i=1}^{k}(P_i - P_{i-1})x_i = \sum_{i=1}^{k} P_i x_i - \sum_{i=1}^{k} P_{i-1}x_i = \\ P_k x_k + \sum_{i=1}^{k} P_{i-1}x_{i-1} - \sum_{i=1}^{k} P_{i-1}x_i = P_k x_k - \sum_{i=1}^{k} P_{i-1}(x_i - x_{i-1}). \end{aligned} \tag{46}$$

Therefore

$$\begin{aligned} P_k(x_k - \tilde{x}_k) =& \sum_{i=1}^{k} P_{i-1}(x_i - x_{i-1}) = \sum_{i=1}^{k}(x_i - x_{i-1}) - \sum_{i=1}^{k}(1 - P_{i-1})(x_i - x_{i-1}) \\ =& x_k - \sum_{i=1}^{k}(1 - P_{i-1})(x_i - x_{i-1}). \end{aligned} \tag{47}$$

Now use condition (43), we obtain

$$P_k(x_k - \tilde{x}_k) = x_k - \sum_{i=1}^{k}(\hat{x}_i - \hat{x}_{i-1}) = x_k - \hat{x}_k, \quad k \geq 1. \tag{48}$$

Thus there holds

$$\hat{x}_k - \tilde{x}_k = x_k - P_k(x_k - \tilde{x}_k) - \tilde{x}_k = (1 - P_t)\,(x_k - \tilde{x}_k), \quad k \geq 1. \tag{49}$$

One can directly verify that the above equations also holds for $k = 0$, which concludes our proof. ■

Lemma 5 *Suppose $x \in \mathbb{R}$. $f(x)$ is α-strongly convex and β-smooth, $0 < \alpha \leq \beta$. Suppose $f(x)$ is lower bounded, then $x_* = \operatorname{argmin}_{x\in\mathbb{R}} f(x)$ exists and is unique. Consider GD with learning rate $0 < \eta < \frac{1}{\beta}$, the optimization path $\{x_k\}_{k=0}^{+\infty}$ is given by*

$$x_{k+1} = x_k - \eta \nabla f(x_k). \tag{50}$$

If $x_0 < x_*$, then we have

1. *For all $k > 0$, $x_k \in (x_0, x_*)$.*
2. *For all $x \in (x_0, x_*)$, we have $\beta(x - x_*) \leq \nabla f(x) \leq \alpha(x - x_*)$.*
3. *For all $x \in (x_0, x_*)$, we have $\alpha(x - x_0) + \nabla f(x_0) \leq \nabla f(x) \leq \beta(x - x_0) + \nabla f(x_0)$.*

Similarly if $x_0 > x_*$, then we have

1. *For all $k > 0$, $x_k \in (x_*, x_0)$.*
2. *For all $x \in (x_*, x_0)$, we have $\alpha(x - x_*) \leq \nabla f(x) \leq \beta(x - x_*)$.*
3. *For all $x \in (x_*, x_0)$, we have $\beta(x - x_0) + \nabla f(x_0) \leq \nabla f(x) \leq \alpha(x - x_0) + \nabla f(x_0)$.*

Proof We only prove Lemma 5 in case of $x_0 < x_*$. The other case is true in a similar manner.

To prove the first conclusion we only need to show that $x_0 < x_1 < x_*$, then recursively we obtain $x_0 < x_1 < \cdots < x_k < x_*$.

Note that $\nabla f(x_*) = 0$. Since $f(x)$ is α-strongly convex and β-smooth, we have [23]

$$\alpha(x - y)^2 \leq \big(\nabla f(x) - \nabla f(y)\big)(x - y) \leq \beta(x - y)^2. \tag{51}$$

Thus $\alpha(x_* - x_0)^2 \leq -\nabla f(x_0)(x_* - x_0) \leq \beta(x_* - x_0)^2$. Now by the assumption that $x_0 < x_*$, we obtain $0 < \alpha(x_* - x_0) \leq -\nabla f(x_0) \leq \beta(x_* - x_0)$. Hence

$$\begin{aligned} &x_1 = x_0 - \eta \nabla f(x_0) > x_0 \\ &x_1 = x_0 - \eta \nabla f(x_0) < x_0 + \eta\beta(x_* - x_0) < x_0 + x_* - x_0 < x_*. \end{aligned} \tag{52}$$

To prove the second conclusion, recall that $\alpha(x_* - x)^2 \leq -\nabla f(x)(x_* - x) \leq \beta(x_* - x)^2$, thus for $x \in (x_0, x_*)$, we obtain $\alpha(x_* - x) \leq -\nabla f(x) \leq \beta(x_* - x)$.

As for the third conclusion, since $\alpha(x - x_0)^2 \leq (\nabla f(x) - \nabla f(x_0))(x - x_0) \leq \beta(x - x_0)^2$, thus for $x \in (x_0, x_*)$, we obtain $\alpha(x - x_0) + \nabla f(x_0) \leq \nabla f(x) \leq \beta(x - x_0) + \nabla f(x_0)$. which completes our proof. ■

Appendix C. Missing proofs in main text

C.1. Proof of Theorem 1

Proof First, provide $0 < \eta < \frac{1}{\beta} < \frac{1}{\alpha}$ and $\gamma = \frac{1}{\frac{1}{\eta}+\lambda}$, we have

$$\frac{\eta\alpha}{\alpha + \lambda} = \frac{1}{\frac{1}{\eta} + \frac{\lambda}{\eta\alpha}} < \frac{1}{\frac{1}{\eta} + \lambda} = \gamma < \frac{1}{\beta + \lambda} \leq \frac{1}{\alpha + \lambda}. \tag{53}$$

Therefore $0 < \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} < 1$, and

$$P_k = 1 - \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad p_k = P_k - P_{k-1}, \tag{54}$$

is a well defined weighting scheme, i.e., P_k is non-negative, non-decreasing and $\lim_{k\to\infty} P_k = 1$.

Let $z_k = \mathbb{E}[w_{k+1}] - \mathbb{E}[w_k]$, $\hat{z}_k = \mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k]$, we first show that

$$(1 - P_k)z_k = \hat{z}_k, \quad k \geq 0. \tag{55}$$

Then according to Lemma 4, we prove the first conclusion in Theorem 1.

The solution of z_k Remember that w_k iterates as

$$w_{k+1} = v_k - \eta(x_{k+1}x_{k+1}^T v_k - x_{k+1}y_{k+1}), \quad v_k = w_k + \tau(w_k - w_{k-1}), \quad w_0 = w_1 = 0, \tag{56}$$

where $\tau = \frac{1-\sqrt{\eta\alpha}}{1+\sqrt{\eta\alpha}}$. Taking expectation with respect to the random mini-batch sampling procedure, since $\mathbb{E}[x_{k+1}x_{k+1}^T] = \frac{1}{n}\sum_{i=1}^n x_i x_i^T = \Sigma$, $\mathbb{E}[x_{k+1}y_{k+1}] = \frac{1}{n}\sum_{i=1}^n x_i y_i = a$, we have

$$\mathbb{E}[w_{k+1}] = \mathbb{E}[v_k] - \eta(\Sigma\mathbb{E}[v_k] - a), \quad \mathbb{E}[v_k] = \mathbb{E}[w_k] + \tau(\mathbb{E}[w_k] - \mathbb{E}[w_{k-1}]), \quad \mathbb{E}[w_0] = \mathbb{E}[w_1] = 0. \tag{57}$$

Eliminate $\mathbb{E}[v_k]$ we obtain

$$\mathbb{E}[w_{k+1}] - (1+\tau)(1-\eta\Sigma)\mathbb{E}[w_k] + \tau(1-\eta\Sigma)\mathbb{E}[w_{k-1}] + \eta a = 0, \quad \mathbb{E}[w_0] = \mathbb{E}[w_1] = 0. \tag{58}$$

Thus $z_k = \mathbb{E}[w_{k+1}] - \mathbb{E}[w_k]$ satisfies

$$z_{k+1} - (1+\tau)(1-\eta\Sigma)z_k + \tau(1-\eta\Sigma)z_{k-1} = 0, \quad z_0 = 0, \quad z_1 = -\eta a. \tag{59}$$

Without loss of generality, let us assume Σ is diagonal in the following. Otherwise consider its eigenvalue decomposition $\Sigma = U\Lambda U^T$, and replace z_k with $U^T z_k$. All of the operators in the following are defined entry-wisely.

Eq. (59) defines a homogeneous linear recurrence relation with constant coefficients, which could be solved in a standard manner. Let

$$A = (1+\tau)(1-\eta\Sigma) = \frac{2(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad B = -\tau(1-\eta\Sigma) = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \tag{60}$$

then the characteristic function of Eq. (59) is

$$r^2 - Ar - B = 0. \tag{61}$$

Since Σ is diagonal, $0 < \eta < \frac{1}{\alpha}$, and α is not greater than the smallest eigenvalue of Σ, we have

$$A^2 + 4B = \frac{4\eta(1-\eta\Sigma)(\alpha-\Sigma)}{(1+\sqrt{\eta\alpha})^2} \leq 0. \tag{62}$$

Thus the characteristic function (61) has two conjugate complex roots r_1 and r_2 (they might be equal). Suppose $r_{1,2} = s \pm ti$. Then the solution of Eq. (59) can be written as

$$z_k = 2(-B)^{\frac{k}{2}}\left(E\cos(\theta k) + F\sin(\theta k)\right), \quad k \geq 0, \tag{63}$$

where E and F are constants decided by initial conditions $z_0 = 0,\ z_1 = -\eta a$, and θ satisfies

$$\cos(\theta) = \frac{s}{\sqrt{s^2+t^2}}, \quad \sin(\theta) = \frac{t}{\sqrt{s^2+t^2}}, \quad r_{1,2} = s \pm ti. \tag{64}$$

Since $2s = r_1 + r_2 = A, s^2 + t^2 = r_1 \dot{r}_2 = -B$, we have

$$\cos(\theta) = \frac{A}{2\sqrt{-B}} = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin(\theta) = \frac{\sqrt{-4B-A^2}}{2\sqrt{-B}} = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}. \tag{65}$$

Because $z_0 = 0, z_1 = -\eta a$, we know that

$$E = 0, \quad 2F = \frac{-\eta a}{(-B)^{\frac{1}{2}}\sin(\theta)}. \tag{66}$$

Thus

$$z_k = \frac{-\eta a}{\sin(\theta)}(-B)^{\frac{k-1}{2}}\sin(\theta k), \quad k \geq 2, \quad z_0 = 0, \quad z_1 = -\eta a. \tag{67}$$

where

$$B = \frac{-(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}, \quad \cos(\theta) = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}}, \quad \sin(\theta) = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}}. \tag{68}$$

One can directly verify that Eq. (67) solves the recurrence relation (59).

The solution of $\hat{z}_k$ Similarly treat the optimization path of the regularized loss, which updates as

$$\hat{w}_{k+1} = \hat{v}_k - \gamma\left((x_{k+1}x_{k+1}^T + \lambda)\hat{v}_k - x_{k+1}y_{k+1}\right), \quad \hat{v}_k = \hat{w}_k + \hat{\tau}(\hat{w}_k - \hat{w}_{k-1}), \quad \hat{w}_0 = \hat{w}_1 = 0, \tag{69}$$

where $\hat{\tau} = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1+\sqrt{\gamma(\alpha+\lambda)}}$. Taking expectation we obtain

$$\mathbb{E}[\hat{w}_{k+1}] = \mathbb{E}[\hat{v}_k] - \gamma((\Sigma+\lambda)\mathbb{E}[\hat{v}_k] - a), \quad \mathbb{E}[\hat{v}_k] = \mathbb{E}[\hat{w}_k] + \hat{\tau}(\mathbb{E}[\hat{w}_k] - \mathbb{E}[\hat{w}_{k-1}]), \tag{70}$$

where $\mathbb{E}[\hat{w}_0] = \mathbb{E}[\hat{w}_1] = 0$. Thus $\hat{z}_k = \mathbb{E}[\hat{w}_{k+1}] - \mathbb{E}[\hat{w}_k]$ satisfies

$$\hat{z}_{k+1} - (1+\hat{\tau})(1-\gamma(\Sigma+\lambda))\hat{z}_k + \hat{\tau}(1-\gamma(\Sigma+\lambda))\hat{z}_{k-1} = 0, \quad \hat{z}_0 = 0, \quad \hat{z}_1 = -\gamma a. \tag{71}$$

Repeat the calculation, we obtain

$$\hat{z}_k = \frac{-\gamma a}{\sin(\hat{\theta})}(-\hat{B})^{\frac{k-1}{2}}\sin(\hat{\theta}k), \quad k \geq 2, \quad \hat{z}_0 = 0, \quad \hat{z}_1 = -\gamma a. \tag{72}$$

where

$$\begin{aligned} \hat{B} &= \frac{-(1-\sqrt{\gamma(\alpha+\lambda)})(1-\gamma(\Sigma+\lambda))}{1+\sqrt{\gamma(\alpha+\lambda)}}, \\ \cos(\hat{\theta}) &= \sqrt{\frac{1-\gamma(\Sigma+\lambda)}{1-\gamma(\alpha+\lambda)}}, \quad \sin(\hat{\theta}) = \sqrt{\frac{\gamma(\Sigma-\alpha)}{1-\gamma(\alpha+\lambda)}}. \end{aligned} \tag{73}$$

Verify Eq. (55) in Lemma 4 First we show that if $1-\lambda\gamma = \frac{\gamma}{\eta}$, we have $\theta \equiv \hat{\theta} \pmod{2\pi}$. To see this, we only need to verify that $\cos(\hat{\theta}) = \cos(\theta),\ \sin(\hat{\theta}) = \sin(\theta)$:

$$\cos(\hat{\theta}) = \sqrt{\frac{1-\gamma\lambda-\gamma\Sigma}{1-\gamma\lambda-\gamma\alpha}} = \sqrt{\frac{\frac{\gamma}{\eta}-\gamma\Sigma}{\frac{\gamma}{\eta}-\gamma\alpha}} = \sqrt{\frac{1-\eta\Sigma}{1-\eta\alpha}} = \cos(\theta); \tag{74}$$

$$\sin(\hat{\theta}) = \sqrt{\frac{\gamma(\Sigma-\alpha)}{1-\gamma\lambda-\gamma\alpha}} = \sqrt{\frac{\gamma(\Sigma-\alpha)}{\frac{\gamma}{\eta}-\gamma\alpha}} = \sqrt{\frac{\eta(\Sigma-\alpha)}{1-\eta\alpha}} = \sin(\theta). \tag{75}$$

Therefore $\sin(\hat{\theta}) = \sin(\theta),\ \sin(\hat{\theta}k) = \sin(\theta k)$, and we have

$$z_k = \frac{-\eta a}{\sin(\theta)}(-B)^{\frac{k-1}{2}}\sin(\theta k), \quad \hat{z}_k = \frac{-\gamma a}{\sin(\theta)}(-\hat{B})^{\frac{k-1}{2}}\sin(\theta k). \tag{76}$$

Since

$$1-P_k = \frac{\gamma}{\eta}\left(\frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}}\right)^{k-1}, \quad \frac{\gamma}{\eta} = 1-\lambda\gamma, \tag{77}$$

we have

$$\begin{aligned}
&\frac{\eta}{\gamma}(1-P_k)(-B)^{\frac{k-1}{2}} = \left(\frac{(1-\sqrt{\gamma(\alpha+\lambda)})^2}{(1-\sqrt{\eta\alpha})^2}\cdot\frac{(1-\sqrt{\eta\alpha})(1-\eta\Sigma)}{1+\sqrt{\eta\alpha}}\right)^{\frac{k-1}{2}} \\
&= \left(\frac{(1-\sqrt{\gamma(\alpha+\lambda)})^2(1-\eta\Sigma)}{1-\eta\alpha}\right)^{\frac{k-1}{2}} = \left(\frac{(1-\sqrt{\gamma(\alpha+\lambda)})^2(1-\gamma(\Sigma+\lambda))}{1-\gamma(\alpha+\lambda)}\right)^{\frac{k-1}{2}} \\
&= \left(\frac{(1-\sqrt{\gamma(\alpha+\lambda)})(1-\gamma(\Sigma+\lambda))}{1+\sqrt{\gamma(\alpha+\lambda)}}\right)^{\frac{k-1}{2}} = (-\hat{B})^{\frac{k-1}{2}}.
\end{aligned} \tag{78}$$

Thus $(1-P_k)z_k = \hat{z}_k$. And according to Lemma 4, we have

$$\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k] = (1-P_k)\left(\mathbb{E}[w_k] - \mathbb{E}[\tilde{w}_k]\right), \quad k \geq 0. \tag{79}$$

The convergence of $\mathbb{E}[\tilde{w}_k]$ Since $L(w)$ is β-smooth, and the corresponding learning rate $\eta < \frac{1}{\beta}$, $\mathbb{E}[w_k]$ converges [3]. Since $\hat{L}(\hat{w}) = L(\hat{w}) + \frac{\lambda}{2}\|\hat{w}\|_2^2$ is $(\beta+\lambda)$-smooth, and the corresponding learning rate $\gamma \leq \frac{1}{\beta+\lambda}$ (see Eq. (53)), $\mathbb{E}[\hat{w}_k]$ converges [3]. Specifically for linear regression, these can be also verified by noticing that $0 < -B < 1$ because $\eta < \frac{1}{\beta}$ and

$$\sum_{i=1}^{k}|z_i| = \sum_{i=1}^{k}\left|\frac{-\eta a}{\sin(\theta)}(-B)^{\frac{i-1}{2}}\sin(\theta i)\right| \leq \sum_{i=1}^{k}\left|\frac{-\eta a}{\sin(\theta)}(-B)^{\frac{i-1}{2}}\right| < +\infty, \tag{80}$$

i.e., the right hand side of the above series converge, which implies that $\mathbb{E}[w_k] = \sum_{i=1}^{k} z_i$ converges absolutely, hence it converges. In a same manner $\mathbb{E}[\hat{w}_k]$ converges.

Thus there exist constants M and K such that for all $k > K, \left\|\mathbb{E}[w_k]\right\|_2 \leq M, \left\|\mathbb{E}[\hat{w}_k]\right\|_2 \leq M$. Hence

$$\left\|\mathbb{E}[\hat{w}_k] - \mathbb{E}[\tilde{w}_k]\right\|_2 = (1-P_k)\left\|\mathbb{E}[w_k] - \mathbb{E}[\hat{w}_k]\right\|_1 \leq \frac{\gamma}{\eta}C^{k-1}\cdot 2M = \mathcal{O}(C^k). \tag{81}$$

Note that $C = \frac{1-\sqrt{\gamma(\alpha+\lambda)}}{1-\sqrt{\eta\alpha}} \in (0,1)$, thus by taking limitation in both sides we obtain

$$\lim_{k\to\infty} \mathbb{E}[\tilde{w}_k] = \lim_{k\to\infty} \mathbb{E}[\hat{w}_k], \tag{82}$$

which concludes our proof. ■

C.2. Proof of Theorem 2

Proof We will prove a stronger version of Theorem 2 by showing the conclusions in Theorem 2 hold along any 1-dim direction $v_1 \in \mathbb{R}^d$. Concisely, given a unit vector $v_1 \in \mathbb{R}^d$, we can extend it to a group of orthogonal basis, $v_1, v_2, \ldots, v_d$. For $w \in \mathbb{R}^d$, we denote its decomposition as

$$w = w^{(1)}v_1 + w^{(2)}v_2 + \cdots + w^{(d)}v_d, \quad w^{(i)} \in \mathbb{R}. \tag{83}$$

Define $h(w^{(1)}) = L(w) = L(w^{(1)}v_1 + \cdots + w^{(d)}v_d)$, then $\nabla h(w^{(1)}) = v_1^T \nabla L(w)$. Now for one step of GD,

$$w_{k+1} = w_k - \eta \nabla L(w_k), \tag{84}$$

by multiplying v_1 in both sides, we obtain

$$w_{k+1}^{(1)} = v_1^T w_{k+1} = v_1^T w_k - \eta v_1^T \nabla L(w_k) = w_k^{(1)} - \eta \nabla h(w_k^{(1)}). \tag{85}$$

We turn to study GD along direction v_1 by analyzing Eq. (85).

Firstly $h(w^{(1)})$ is α-strongly convex, β-smooth and lower bounded since $L(w)$ is α-strongly convex, β-smooth, and lower bounded. Let w_* be the unique minimal of $L(w)$, then $w_*^{(1)} = v_1^T w_*$ is the minimal of $h(w^{(1)})$. With out losing generality, assume

$$w_*^{(1)} > 0 = w_0^{(1)}. \tag{86}$$

Then by Lemma 5, we know the optimization path of Eq. (85) lies between $(w_*^{(1)}, 0)$, and for any $v \in (w_*^{(1)}, 0)$, we have

$$\alpha v - b \le \nabla h(v) \le \beta v - b, \quad b = -\nabla L(0). \tag{87}$$

Thus for Eq. (85) we have

$$\begin{aligned} w_{k+1}^{(1)} - w_k^{(1)} &= -\eta \nabla h(w_k^{(1)}) \le -\eta(\alpha w_k^{(1)} - b), \\ w_{k+1}^{(1)} - w_k^{(1)} &= -\eta \nabla h(w_k^{(1)}) \ge -\eta(\beta w_k^{(1)} - b). \end{aligned} \tag{88}$$

Define the following dynamics:

$$u_{k+1}^{(1)} - u_k^{(1)} = -\eta(\alpha u_k^{(1)} - b), \quad v_{k+1}^{(1)} - v_k^{(1)} = -\eta(\beta v_k^{(1)} - b), \quad u_0^{(1)} = v_0^{(1)} = 0. \tag{89}$$

By the discrete Gronwall's inequality [4], we have

$$v_k^{(1)} \le w_k^{(1)} \le u_k^{(1)}. \tag{90}$$

Furthermore, $u_k^{(1)}$ and $v_k^{(1)}$ satisfy two first order recurrence relations respectively, thus they are solved by

$$u_k^{(1)} = \eta \sum_{i=1}^{k} (1-\eta\alpha)^{i-1} b, \quad v_k^{(1)} = \eta \sum_{i=1}^{k} (1-\eta\beta)^{i-1} b. \tag{91}$$

Since $\eta < \frac{1}{\beta} \le \frac{1}{\alpha}$, u_k and v_k converge, and $\lim_{k\to+\infty} u_k = \lim_{k\to+\infty} v_k = w_*^{(1)}$. $w_k^{(1)}$ also converges since $L(w)$ is β-smooth convex and $\eta < \frac{1}{\beta}$.

In a same way, for the regularized path,

$$\hat{w}_{k+1,\lambda}^{(1)} = \hat{w}_{k,\lambda}^{(1)} - \gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}), \quad \hat{w}_{0,\lambda}^{(1)} = 0, \tag{92}$$

we have

$$\begin{aligned} \hat{w}_{k+1,\lambda}^{(1)} - \hat{w}_{k,\lambda}^{(1)} &= -\gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}) \le -\gamma\left((\alpha+\lambda)\hat{w}_{k,\lambda}^{(1)} - b\right), \\ \hat{w}_{k+1,\lambda}^{(1)} - \hat{w}_{k,\lambda}^{(1)} &= -\gamma(\nabla h(\hat{w}_{k,\lambda}^{(1)}) + \lambda \hat{w}_{k,\lambda}^{(1)}) \ge -\gamma\left((\beta+\lambda)\hat{w}_{k,\lambda}^{(1)} - b\right). \end{aligned} \tag{93}$$

Consider the following dynamics:

$$\hat{u}_{k+1,\lambda}^{(1)} - \hat{u}_{k,\lambda}^{(1)} = -\gamma((\alpha+\lambda)\hat{u}_{k,\lambda}^{(1)} - b), \quad \hat{v}_{k+1,\lambda}^{(1)} - \hat{v}_{k,\lambda}^{(1)} = -\gamma((\beta+\lambda)\hat{v}_{k,\lambda}^{(1)} - b), \tag{94}$$

where $\hat{u}_{0,\lambda}^{(1)} = \hat{v}_{0,\lambda}^{(1)} = 0$. Then by the discrete Gronwall's inequality and the solution of the first order recurrence relation we obtain

$$\hat{v}_{k,\lambda}^{(1)} \le \hat{w}_{k,\lambda}^{(1)} \le \hat{u}_{k,\lambda}^{(1)}, \quad \hat{u}_{k,\lambda}^{(1)} = \gamma \sum_{i=1}^{k} (1-\gamma(\alpha+\lambda))^{i-1} b, \quad \hat{v}_{k,\lambda}^{(1)} = \gamma \sum_{i=1}^{k} (1-\gamma(\beta+\lambda))^{i-1} b. \tag{95}$$

Now we turn to bound the iterate averaged solution. Consider

$$\lambda_1 = \frac{1}{\gamma} - \frac{1}{\eta} + \beta - \alpha, \quad \lambda_2 = \frac{1}{\gamma} - \frac{1}{\eta} + \alpha - \beta, \tag{96}$$

since $\beta \ge \alpha$ and $0 < \gamma < \frac{1}{\beta-\alpha+1/\eta}$ we know $\lambda_1 \ge \lambda_2 > 0$. Notice that $0 < \gamma(\alpha+\lambda_2) \le \{\gamma(\alpha+\lambda_1),\ \gamma(\beta+\lambda_2)\} \le \gamma(\beta+\lambda_1) = 1 - \gamma(-\frac{1}{\eta} + 2\beta - \alpha) < 1$, where the last inequality is because $\eta > \frac{1}{2\beta-\alpha}$. Thus $\hat{u}_{k,\lambda_1}^{(1)}$, $\hat{u}_{k,\lambda_2}^{(1)}$, $\hat{v}_{k,\lambda_1}^{(1)}$, $\hat{v}_{k,\lambda_2}^{(1)}$ converge. Further $\hat{w}_{k,\lambda_1}$ and $\hat{w}_{k,\lambda_2}$ also converge since $\gamma < \frac{1}{\beta+\lambda_1} \le \frac{1}{\beta+\lambda_2}$ and the corresponding regularized losses are $(\beta+\lambda_1)$ and $(\beta+\lambda_2)$-smooth, respectively.

Next let us consider the weight scheme $P_k = 1 - \left(\frac{\gamma}{\eta}\right)^{k+1}$, which is well defined since $0 < \gamma < \frac{1}{\beta-\alpha+1/\eta} \le \eta$.

One can directly verify that $\tilde{u}_k^{(1)} = \frac{1}{P_k}\sum_{i=1}^{k} p_i u_i^{(1)}$, $\tilde{v}_k^{(1)} = \frac{1}{P_k}\sum_{i=1}^{k} p_i v_i^{(1)}$ converge, and

$$(1-P_k)(u_{k+1}^{(1)} - u_k^{(1)}) = \hat{v}_{k+1,\lambda_2}^{(1)} - \hat{v}_{k,\lambda_2}^{(1)}, \quad (1-P_k)(v_{k+1}^{(1)} - v_k^{(1)}) = \hat{u}_{k+1,\lambda_1}^{(1)} - \hat{u}_{k,\lambda_1}^{(1)}. \tag{97}$$

Thus according to Lemma 4 we have

$$P_k(u_k^{(1)} - \tilde{u}_k^{(1)}) = u_k^{(1)} - \hat{v}_{k,\lambda_2}^{(1)}, \quad P_k(v_k^{(1)} - \tilde{v}_k^{(1)}) = v_k^{(1)} - \hat{u}_{k,\lambda_1}^{(1)}. \tag{98}$$

Therefore

$$\begin{aligned}\tilde{w}_k^{(1)} - \hat{w}_{k,\lambda_2}^{(1)} \le \tilde{u}_k^{(1)} - \hat{v}_{k,\lambda_2}^{(1)} = \tilde{u}_k^{(1)} - u_k^{(1)} + P_k(u_k^{(1)} - \tilde{u}_k^{(1)}) = (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}),\\ \tilde{w}_k^{(1)} - \hat{w}_{k,\lambda_1}^{(1)} \ge \tilde{v}_k^{(1)} - \hat{u}_{k,\lambda_1}^{(1)} = \tilde{v}_k^{(1)} - v_k^{(1)} + P_k(v_k^{(1)} - \tilde{v}_k^{(1)}) = (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}),\end{aligned} \tag{99}$$

which implies that

$$\hat{w}_{k,\lambda_1}^{(1)} + (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}) \le \tilde{w}_k^{(1)} \le \hat{w}_{k,\lambda_2}^{(1)} + (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}). \tag{100}$$

This completes the proof of the first conclusion in Theorem 2.

Note that $u_k^{(1)},\ \tilde{u}_k^{(1)},\ v_k^{(1)},\ \tilde{v}_k^{(1)},\ \hat{w}_{k,\lambda_1}^{(1)},\ \hat{w}_{k,\lambda_2}^{(1)}$ converge, therefore there is a constant M bound their ℓ_2-norm. Define $m_k^{(1)} = (\hat{w}_{k,\lambda_2}^{(1)} + \hat{w}_{k,\lambda_1}^{(1)})/2$, $d_k^{(1)} = (\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{k,\lambda_1}^{(1)})/2$. Recall that $\hat{w}_{k,\lambda_1}^{(1)}$ are the GD optimization path of a $(\alpha+\lambda_1)$-strongly convex and $(\beta+\lambda_1)$-smooth loss, thus $\hat{w}_{k,\lambda_1}^{(1)}$ converges in rate $\mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right)$. Similarly $\hat{w}_{k,\lambda_2}^{(1)}$ converges in rate $\mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right)$. Thus by triangle inequality we have

$$\begin{aligned}\left\|m_k^{(1)} - m^{(1)}\right\|_2 \le& \frac{1}{2}\left\|\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{\infty,\lambda_2}^{(1)}\right\|_2 + \frac{1}{2}\left\|\hat{w}_{k,\lambda_1}^{(1)} - \hat{w}_{\infty,\lambda_1}^{(1)}\right\|_2\\ \le& \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right).\\ \left\|d_k^{(1)} - d^{(1)}\right\|_2 \le& \frac{1}{2}\left\|\hat{w}_{k,\lambda_2}^{(1)} - \hat{w}_{\infty,\lambda_2}^{(1)}\right\|_2 + \frac{1}{2}\left\|\hat{w}_{k,\lambda_1}^{(1)} - \hat{w}_{\infty,\lambda_1}^{(1)}\right\|_2\\ \le& \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right).\end{aligned} \tag{101}$$

By Eq. (100) we obtain

$$\begin{aligned}\tilde{w}_k^{(1)} - m_k^{(1)} \le& d_k^{(1)} + (1-P_k)(\tilde{u}_k^{(1)} - u_k^{(1)}) \le d_k^{(1)} + 2M\left(\frac{\gamma}{\eta}\right)^{k+1}\\ =& d^{(1)} + d^{(1)} - d_k^{(1)} + \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right)\\ \le& d^{(1)} + \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) + \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right) + \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right).\\ \tilde{w}_k^{(1)} - m_k^{(1)} \ge& d_k^{(1)} + (1-P_k)(\tilde{v}_k^{(1)} - v_k^{(1)}) \ge d_k^{(1)} - 2M\left(\frac{\gamma}{\eta}\right)^{k+1}\\ =& d^{(1)} + d^{(1)} - d_k^{(1)} - \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right)\\ \ge& d^{(1)} - \mathcal{O}\left((1-\gamma(\alpha+\lambda_1))^k\right) - \mathcal{O}\left((1-\gamma(\alpha+\lambda_2))^k\right) - \mathcal{O}\left(\left(\frac{\gamma}{\eta}\right)^k\right).\end{aligned} \tag{102}$$

Thus

$$\left\|\tilde{w}_k^{(1)} - m_k^{(1)}\right\|_2 \le \mathcal{O}(C^k), \quad C = \max\{(1-\gamma(\alpha+\lambda_1), (1-\gamma(\alpha+\lambda_2), \frac{\gamma}{\eta}\}. \tag{103}$$

Therefore

$$\left\|\tilde{w}_k^{(1)} - m^{(1)}\right\|_2 \leq \left\|\tilde{w}_k^{(1)} - m_k^{(1)}\right\|_2 + \left\|m_k^{(1)} - m^{(1)}\right\|_2 \leq \mathcal{O}(C^k), \tag{104}$$

which completes our proof.

■