High Dimensional First Order Mini-Batch Algorithms on Quadratic Problems

author names withheld

Under Review for OPT 2024

Abstract

We analyze the dynamics of general mini-batch first order algorithms on the ℓ_2 regularized least squares problem when the number of samples and dimensions are large. This includes stochastic gradient descent (SGD), stochastic Nesterov (convex/strongly convex), and stochastic momentum. In this setting, we show that the dynamics of these algorithms concentrate to a deterministic discrete Volterra equation Ψ in the high-dimensional limit. In turn, we show that we can use Ψ to capture the behaviour of general mini-batch first order algorithm under *any* quadratic statistics $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ (see Definition 6), including but not limited to: training loss, excess risk for empirical risk minimization (in-distribution and generalization error.

1. Main

The structure of an optimization problem plays an important role in designing efficient algorithms. A common structure, motivated by empirical risk minimization (ERM), is the finite sum, that is, an optimization problem of the form

$$\min_{x\in\mathbb{R}^d} \left\{ f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}) \right\}, \tag{1}$$

where the functions $f_i : \mathbb{R}^d \to \mathbb{R}$. Much has been written about the complexity for solving (1) under various assumptions on f, such as smoothness and convexity [4, 6–11, 14, 15, 17–19, 23, 25, 31, 32].

Motivated in part by the rise in machine learning, weakening these assumptions, as many problems of interest are nonsmooth and nonconvex, have dominated current optimization research. This has led to tight upper bounds on complexity that match the theoretic lower bounds for very general finite-sum problems [3, 12, 13, 26], and yet in spite of this, there exists an enormous gap between these theoretical guarantees and observed performance on machine learning problems. Indeed, even in the smooth, convex setting, there is a missing component in our understanding of finite sum problems in machine learning. One possibility is simply the size of the finite sums. An overarching trend in machine learning is to scale up the problem size, measured both by model complexity (parameters) and data set size. In short, machine learning problems are *high-dimensional*.

Beyond high-dimensionality, machine learning problems also tend to be stochastic: the data are random, the learning algorithms are random, and the model initialization is random. This trifecta of randomness combined with high-dimensionality may account for this missing structure from optimization theory for machine learning.

This article develops a framework for first-order stochastic methods that incorporates high-dimensionality for analyzing all first-order stochastic learning algorithms on an ℓ^2-regularized least squares problem. The framework, first proposed for analyzing stochastic gradient descent (SGD) in the "small batch regime" [29], imports mathematics ideas commonly used in random matrix theory.

Notation. Throughout the article, we adhere to the following. We write vectors in lowercase boldface ($\boldsymbol{x}$) and matrices in uppercase boldface ($\boldsymbol{H}$). The norm $\|\boldsymbol{x}\|_2^2 = \boldsymbol{x}^T\boldsymbol{x}$ gives the usual Euclidean 2-norm and $\|\boldsymbol{H}\| =$ maximum singular value of $\boldsymbol{H}$ is the usual operator 2-norm.

1.1. Formal Setup and Assumptions.

To formalize the analysis, we define the *ℓ^2-regularized least-squares problem:*

$$\arg\min_{x\in\mathbb{R}^d}\left\{f(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x}-\boldsymbol{b}\|^2 + \frac{\delta}{2}\|\boldsymbol{x}\|^2 = \sum_{i=1}^{n}\underbrace{\frac{1}{2}\left((\boldsymbol{a}_i\boldsymbol{x}-b_i)^2 + \frac{\delta}{n}\|\boldsymbol{x}\|^2\right)}_{\stackrel{\text{def}}{=} f_i(\boldsymbol{x})}\right\}. \tag{2}$$

The fixed parameter $\delta > 0$ controls the regularization strength and it is independent of n and d. We focus on setups where the parameter choices n and d are large, but we do not require that they are proportional. This is captured by Assumption 2.

Assumptions on Data. Moreover our results only gain power when one (and hence both) of these parameters are large. The data matrix $\boldsymbol{A} \in \mathbb{R}^{n\times d}$ and the labels $\boldsymbol{b}$ may be deterministic or random; we formulate our theorems for deterministic $\boldsymbol{A}$ and $\boldsymbol{b}$ in (2) satisfying various assumptions, and in the applications of our theorems to statistical settings, we shall show that the random $\boldsymbol{A}$ and $\boldsymbol{b}$ (2) satisfy these assumptions.

These assumptions are motivated by the case where the augmented matrix $[\boldsymbol{A}\,|\,\boldsymbol{b}]$ has rows that are independent and sampled from some common distribution. We also note that the problem (2) is homogeneous, in that if we divide all of $\boldsymbol{A}$, $\boldsymbol{b}$, and $\sqrt{\delta}$ by any desired scalar, we produce an equivalent optimization problem. As such, we adopt the following convention without loss of generality.

Assumption 1 (Data–target normalization) *There is a constant $C > 0$ independent of d and n such that the spectral norm of $\boldsymbol{A}$ is bounded by C and the target vector $\boldsymbol{b} \in \mathbb{R}^n$ is normalized so that $\|\boldsymbol{b}\|^2 \le C$.*

More importantly, we also assume that the data and targets resemble typical unstructured high-dimensional random matrices. One of the principal qualitative properties of high-dimensional random matrices is the *delocalization of their eigenvectors*, which refers to the statistical similarity of the eigenvectors to uniform random elements from the Euclidean sphere. The precise mathematical description of this assumption is most easily given in terms of resolvent bounds. The resolvent $R(z;\boldsymbol{M})$ of a matrix $\boldsymbol{M} \in \mathbb{R}^{d\times d}$ is

$$R(z;\boldsymbol{M}) = (\boldsymbol{M} - z\mathbf{I}_d)^{-1} \quad \text{for } z \in \mathbb{C}.$$

In terms of the resolvent, this is formally captured by Assumption 3.

Definition 1 ((deterministic) Gradient-based method) A deterministic optimization algorithm is called a *gradient-based method* if each update of the algorithm can be written as a linear combination of the previous iterate and previous gradient. In other words, if every update is of the form

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \nabla f(\boldsymbol{x}_j), \tag{3}$$

for some scalar values $c_{k,j}$.

Examples of gradient-based methods include momentum methods [30], accelerated methods [27], and gradient descent. Given any such (deterministic) gradient-based method, we can construct a stochastic (mini-batch) version by approximating the full gradient $\nabla f(\boldsymbol{x}_i)$ with a stochastic version. In particular, we choose a batch $B_k \subset \{1, 2, \ldots, n\}$ of cardinality β uniformly at random and approximate $\nabla f(\boldsymbol{x}_i) \approx \sum_{i \in B_j} \nabla f_i(\boldsymbol{x}_j)$. We focus on these stochastic algorithms. The resulting mini-batch, gradient based algorithms are considered in this work. Specifically, our goal is to provide the (multi-pass) dynamics for *any* mini-batch version of the gradient-based methods on the ℓ^2-regularized least squares problem (2).

Definition 2 (Mini-batch, gradient-based method) Given a gradient-based algorithm, we define a stochastic optimization algorithm, called a *mini-batch, gradient-based method*, if at each update one generated uniformly at random a batch $B_i \subset \{1, 2, \ldots, n\}$ and the update satisfies,

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k} c_{k,j} \sum_{i \in B_j} \nabla f_i(\boldsymbol{x}_j) \tag{4}$$

for some scalars $c_{k,i}$. Note that the scalars $c_{k,i}$ are general enough to allow for time-dependent learning rates and momentum parameters.

We denote the proportion of the number of samples n relative to the size of the batch B_i taken by the mini-batch gradient-based method as the *batch fraction*,

$$\zeta \stackrel{\text{def}}{=} \frac{|B_i|}{n} = \frac{\beta}{n}. \tag{5}$$

Next, we introduce a natural comparison to the mini-batch algorithm whose iterates evolves as

$$\boldsymbol{y}_{k+1} = \boldsymbol{x}_0 + \zeta \sum_{j=0}^{k} c_{k,j} \nabla f(\boldsymbol{y}_k), \quad k \geq 0, \tag{6}$$

where the $\boldsymbol{y}_k$ iterates are initialized also at the point $\boldsymbol{x}_0$ (i.e., $\boldsymbol{y}_0 = \boldsymbol{x}_0$). The $\boldsymbol{y}_k$ iterates are the corresponding full-batch gradient-based method whose coefficients $c_{k,j}$ are scaled by the batch fraction ζ.

Assumptions on $\boldsymbol{x}_0$. As for the initialization $\boldsymbol{x}_0$, we need to suppose that it does not interact too strongly with the *right* singular-vectors of $\boldsymbol{A}$. Formally, this is captured by Assumption 4. Note that, as a simple but common initialization choice, Assumption 4 is surely satisfied for $\boldsymbol{x}_0 = \boldsymbol{0}$. In principle, this assumption is general enough to allow $\boldsymbol{x}_0$ which are correlated with $\boldsymbol{A}$ in a nontrival way. However one common (nonzero) initialization scheme is to choose $\boldsymbol{x}_0$ independent of $(\boldsymbol{A}, \boldsymbol{b})$.

Finally while we are not able to handle all statistics, we will be able to give descriptions for quadratic statistics given by Definition 6.

1.2. Main Theorem and Applications

One of our main results is a (non asymptotic) description of the training and population risk curves of any mini-batch, gradient-based algorithm to deterministic functions whose accuracy improves when the number of samples and features are large. Moreover, our analysis generalizes to generate descriptions of these algorithms under *any* quadratic statistic $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ (see Definition 6). The result provides a unifying theory that holds for any gradient-based, mini-batch method (see Definition 2). To do so, we will refer frequently to the empirical risk

$$\mathcal{L}(\boldsymbol{x}) \stackrel{\text{def}}{=} \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2. \tag{7}$$

In this section, we present the informal version of our main result and defer the formal version to Theorem 7 in the appendix.

Theorem 3 ((Informal) Dynamics of Mini-Batch, Gradient-Based Methods) *Suppose Assumptions 1, 2, 3, and 4 hold on the data matrix $\boldsymbol{A}$, targets $\boldsymbol{b}$, and initialization $\boldsymbol{x}_0$ with $\alpha_0 \in (0, 1/4)$. Let $\mathcal{R}(\cdot)$ be a quadratic statistic satisfying Assumption 5. Consider the iterates $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ generated by a mini-batch, gradient-based algorithm (4) with batch size satisfying $\beta/n = \zeta$ for some $\zeta > 0$. For fixed $T > 0$ there exists $\tilde{C} > 0$ such that for all $c > 0$ there exists $D > 0$ satisfying*

$$\Pr\left[\sup_{0\le t\le T}\left\|\begin{pmatrix}\mathcal{L}(\boldsymbol{x}_t)\\ \mathcal{R}(\boldsymbol{x}_t)\end{pmatrix} - \begin{pmatrix}\Psi(t)\\ \Omega(t)\end{pmatrix}\right\| > C(T)n^{-\tilde{C}}\right] \le Dn^{-c}, \tag{8}$$

where

$$\begin{aligned}\Psi(t) &\stackrel{def}{=} \mathcal{L}(\boldsymbol{y}_t) + \mathcal{K}(\nabla^2\mathcal{L})_{0:t-1} * \Psi_{1:t-1}\\ \Omega(t) &\stackrel{def}{=} \mathcal{R}(\boldsymbol{y}_t) + \mathcal{K}(\nabla^2\mathcal{R})_{0:t-1} * \Psi_{0:t-1}\end{aligned} \tag{9}$$

where $\mathcal{K}(\mathbf{M})_{0:t-1} * \Psi_{1:t-1}$ denotes the convolution between $\mathcal{K}(\mathbf{M})$, a kernel dependent on a matrix $\mathbf{M}$ evaluated at t points, and the previous values $\Psi_{0:t-1} \stackrel{def}{=} \{\Psi(0), \dots, \Psi(t-1)\}$.

Figure 1 illustrates capturing the training dynamics with $\Psi(t)$ under SGD+M [30] and Nesterov acceleration [27]. Figure 3 illustrates $\Omega(t)$ capturing the *in-generalization distribution risk* (see Sect. 1.2.2) under SGD+M.

Interpretation of Main Theorem. In the high-dimensional limit, $\Psi(t)$ captures the dynamics of $\mathcal{L}(\boldsymbol{x}_t)$. In particular, $\Psi(t)$ can be viewed as a summation between two terms, $\mathcal{L}(\boldsymbol{y}_t)$, which represents the *deterministic* behaviour of $\mathcal{L}$ under the mini-batch algorithm (recall $\boldsymbol{y}_t$ are the iterates corresponding full-batch gradient-based method scaled by ζ), and a convolution term, $\mathcal{K} * \Psi$, which represents the noise inherent in the algorithm (i.e., the kernel $\mathcal{K}$ contains the noise information generated by random batch sampling). $\Omega(t)$ captures the dynamics of the quadratic statistic $\mathcal{R}(t)$, in the same manner.

Theorem 7 generalizes the findings of [21] which are restricted to SGD+M on the training loss $\mathcal{L}$ to include *all* mini-batch first order algorithms and *all* quadratic functions $\mathcal{R}(\cdot)$. Specifically, since $\mathcal{L}(\boldsymbol{x}_t)$ and $\mathcal{R}(\boldsymbol{x}_t)$ is captured by $\Psi(t)$ and $\Omega(t)$ in the high-dimensional limit, respectively, we can analyze $\Psi(t)$ and $\Omega(t)$ directly to analyze fundamental properties of any stochastic mini-batch algorithm including but not limited to optimal batch-size, selection of optimal hyperparameters, and convergence regimes. In the next sections, we give some motivating problem setups that illustrate the versatility of our setup and some common statistics.

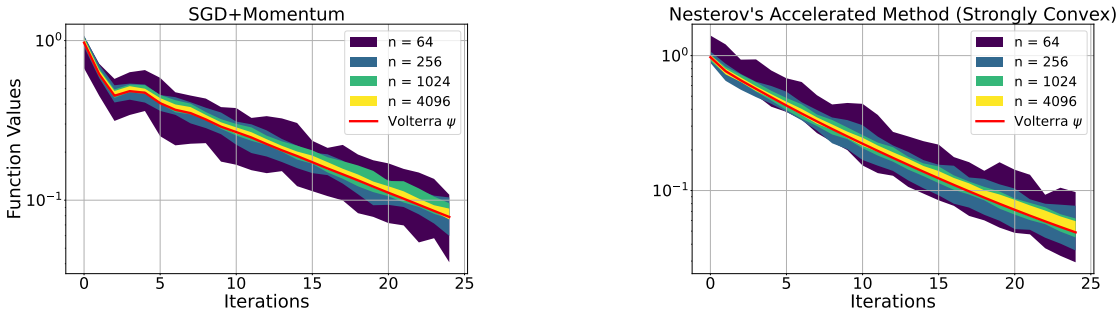

Figure 1: **Concentration of mini-batch algorithms on Gaussian random least squares problem.** The 90th percentile confidence interval of various mini-batch algorithms are plotted (shaded regions) with fixed ratio d/n. The figure demonstrates the loss $\mathcal{L}(\boldsymbol{x}_t)$ concentrates to the Volterra equation $\Psi(t)$ (see Theorem 7). The hyperparameters are fixed at momentum $\Delta = 0.8$, learning rate $\gamma = 0.5$, and batch fraction $\zeta = 0.5$ across the experiments.

1.2.1. TRAINING LOSS

Sample covariance matrices and generative models. A natural assumption under which the data $(\boldsymbol{A}, \boldsymbol{b})$ satisfies Assumptions 1 and 3 is for *sample covariance matrices* and *generative models*. For this, suppose $\boldsymbol{\Sigma} \succeq 0$ is a $d \times d$ matrix with $\text{tr}(\boldsymbol{\Sigma}) = 1$ and $\|\boldsymbol{\Sigma}\| \leq M/\sqrt{d} < \infty$ for some constant $M > 0$. We construct the data (random) matrix $\boldsymbol{A}$ by setting $\boldsymbol{A} = \boldsymbol{Z}\sqrt{\boldsymbol{\Sigma}}$ where $\boldsymbol{Z}$ is an $n \times d$ matrix of independent, mean 0, variance 1 entries with subgaussian norm at most $M < \infty$. We also assume that $n \leq Md$. Finally, suppose that $\boldsymbol{b}$ satisfies the generative model, that is, $\boldsymbol{b} = \boldsymbol{A}\boldsymbol{\beta} + \boldsymbol{\eta}$ for $\boldsymbol{\beta}, \boldsymbol{\eta}$ iid centered subgaussians satisfying $\|\boldsymbol{b}\|^2 = R$ and $\|\boldsymbol{\eta}\| = \tilde{R}\frac{n}{d}$ for some $R, \tilde{R} > 0$. It follows from Lemma 1.3 in [29] that the sample covariance matrix $\boldsymbol{A}$ and generative model $\boldsymbol{b}$ generated this way satisfy Assumptions 1 and 3. Hence under these assumptions, we conclude:

Theorem 4 (Dynamics of training loss with sample covariance matrices and generative models) *Suppose $(\boldsymbol{A}, \boldsymbol{b})$ is a sample covariance matrix and generative model framework. Let $\delta > 0$ and $\boldsymbol{x}_0$ is iid centered subgaussian with $\mathbb{E}\left[\|\boldsymbol{x}_0\|^2\right] = \hat{R}$. Then for some $\varepsilon > 0$, for all $T > 0$, and for all $D > 0$, there is a $C > 0$ such that*

$$\Pr\left(\sup_{0 \leq t \leq T} \left\| \begin{pmatrix} \mathcal{L}(\boldsymbol{x}_t) \\ \frac{1}{2}\|\boldsymbol{x}_t - \boldsymbol{\beta}\|^2 \end{pmatrix} - \begin{pmatrix} \Psi(t) \\ \Omega(t) \end{pmatrix} \right\| \geq n^{-\varepsilon}\right) \leq Cn^{-D},$$

where $\Psi(t)$ and $\Omega(t)$ solves (16) with $\mathcal{R} = \frac{1}{2}\|\cdot - \boldsymbol{\beta}\|^2$.

Figure 3 illustrate capturing the dynamics of training loss of various popular mini-batch algorithms.

1.2.2. EXCESS RISK FOR ERM IN LINEAR REGRESSION

One standard linear regression setup supposes that $\boldsymbol{A}$ is generated by taking n independent d-dimensional samples from a centered distribution $\mathcal{D}_f$. Here a distribution is standardized if the

distribution has mean 0 and expected sample-norm-squared equal to 1. Let the matrix $\mathbf{\Sigma}_f$ be the $d \times d$ feature covariance matrix of $\mathcal{D}_f$,

$$\mathbf{\Sigma}_f \stackrel{\text{def}}{=} \mathbb{E}\left[\boldsymbol{a}\boldsymbol{a}^T\right], \qquad \text{where } \boldsymbol{a} \sim \mathcal{D}_f. \tag{10}$$

Now suppose there is a linear ("ground truth" or "signal") function $\beta : \mathbb{R}^d \to \mathbb{R}$ which, for simplicity, has $\beta(0) = 0$. Since we are doing linear regression, we identify β with a vector using the representation $\boldsymbol{a} \mapsto \boldsymbol{\beta}^T\boldsymbol{a}$. We suppose that our data $(\boldsymbol{a}, b)$ is draw from a distribution $\mathcal{D}$ on $\mathbb{R}^d \times \mathbb{R}$, with the property that

$$\mathbb{E}\left[b \,|\, \boldsymbol{a}\right] = \boldsymbol{\beta}^T\boldsymbol{a}, \quad \text{where} \quad (\boldsymbol{a}, b) \sim \mathcal{D} \quad \text{and} \quad \boldsymbol{a} \sim \mathcal{D}_f.$$

Hence we suppose that $[\boldsymbol{A} \,|\, \boldsymbol{b}]$ is a $\mathbb{R}^{n\times d} \times \mathbb{R}^{n\times 1}$ matrix on independent samples from $\mathcal{D}$. The iterate from training $\boldsymbol{x}_t$ represents an estimate of $\boldsymbol{\beta}$, and the population risk is

$$\mathcal{R}(\boldsymbol{x}_t) \stackrel{\text{def}}{=} \frac{1}{2}\mathbb{E}\left[(b - \boldsymbol{x}_t^T\boldsymbol{a})^2 \,|\, \boldsymbol{x}_t\right],$$

where $(\boldsymbol{a}, b) \sim \mathcal{D}$ and independent sample from generating the estimate $\boldsymbol{x}_t$. In this setting, the estimate $\boldsymbol{x}_t$ is generated from samples distributed as $\mathcal{D}$ *and* the population risk is evaluated with a sample from the same distribution. We call this *in-distribution*. When the estimate is generated from a different distribution than the evaluation of the population risk we call it *out-of-distribution*.

We can evaluate the population risk in terms of the feature covariance matrix $\mathbf{\Sigma}_f$ and the noise $\eta^2 \stackrel{\text{def}}{=} \mathbb{E}\left[(b - \boldsymbol{\beta}^T\boldsymbol{a})^2\right]$ to give

$$\mathcal{R}(\boldsymbol{x}_t) = \frac{1}{2}\eta^2 + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{x}_t)^T\mathbf{\Sigma}_f(\boldsymbol{\beta} - \boldsymbol{x}_t). \tag{11}$$

Here the sequence of iterates $\{\boldsymbol{x}_t\}_{t\geq 0}$ is generated from the mini-batch, gradient-based algorithm applied to the ℓ^2-regularized least-squares problem (2).

In the case that $(\boldsymbol{a}, b)$ are jointly Gaussian, we may represent

$$\boldsymbol{a} = \mathbf{\Sigma}_f^{1/2}\boldsymbol{z}, \quad b = \boldsymbol{\beta}^T\boldsymbol{a} + \eta w, \quad \text{where } (\boldsymbol{z}, w) \sim N(0, \mathbf{I}_d \oplus 1).$$

This is to say that $(\boldsymbol{a}, b)$ come from a sample covariance with covariance $\mathbf{\Sigma}_f$ and generative model with signal $\boldsymbol{\beta}$.

If $\mathcal{D}$ satisfies the the conditions of sample covariance matrices (with $\mathbf{\Sigma} = \mathbf{\Sigma}_f$), then the population risk $\mathcal{R}$ is well-approximated by Ω:

Theorem 5 *Suppose $(\boldsymbol{A}, \boldsymbol{b})$ satisfy the sample covariance and generative model framework. Let $\delta > 0$ and $\boldsymbol{x}_0$ an centered iid subgaussian vector with $\mathbb{E}\left[\|\boldsymbol{x}_0\|^2\right] = \hat{R}$. Then for some $\varepsilon > 0$, for all $T > 0$, and for all $D > 0$, there is a $C > 0$ such that*

$$\Pr\left(\sup_{0\leq t\leq T}\left\|\begin{pmatrix}\mathcal{L}(\boldsymbol{x}_t)\\ \mathcal{R}(\boldsymbol{x}_t)\end{pmatrix} - \begin{pmatrix}\Psi(t)\\ \Omega(t)\end{pmatrix}\right\| \geq n^{-\varepsilon}\right) \leq Cn^{-D},$$

where $\Psi(t)$ and $\Omega(t)$ solves (16) with $\mathcal{R}(\boldsymbol{x})$ the population risk defined in (11).

When the set-up is "in-distribution" and $\boldsymbol{b}$ follows the generative model, then $\eta^2 = \frac{\tilde{R}}{d}$. For "out-of-distribution", this is not the case. The noise $\eta^2 \neq \frac{\tilde{R}}{d}$ as η represents the population noise. For the sake of space, we will not include the out-of-distribution case, but note that it can be readily addressed. One can also apply this set-up to *Random features* with the set-up based upon [2, 22].

References

[1] R. Adamczak. A note on the Hanson-Wright inequality for random vectors with dependencies. *Electronic Communications in Probability*, 20(none):1 – 13, 2015.

[2] B. Adlam and J. Pennington. The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pages 74–84, 2020.

[3] Y. Arjevani, Y. Carmon, and J.C. Duchi. Lower bounds for non-convex stochastic optimization. *Math. Program.*, 155, 2022. doi: 10.1007/s10107-022-01822-7. URL https://doi.org/10.1007/s10107-022-01822-7.

[4] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 26, 2013.

[5] R. Bardenet and Odalric-Ambrym M. Concentration inequalities for sampling without replacement. *Bernoulli*, 21(3):1361 – 1385, 2015.

[6] D. Bertsekas. A new class of incremental gradient methods for least squares problems. *SIAM J. Optim.*, 7(4):913–926, 1997.

[7] D. Bertsekas and J. Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM J. Optim.*, 10(3):627–642, 2000.

[8] L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[9] D. Davis and D. Drusvyatskiy. Stochastic model-based minimization of weakly convex functions. *SIAM J. Optim.*, 29(1):207–239, 2019. ISSN 1052-6234. doi: 10.1137/18M1178244. URL https://doi.org/10.1137/18M1178244.

[10] Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, 2014.

[11] A. Defossez and F. Bach. Averaged Least-Mean-Squares: Bias-Variance Trade-offs and Optimal Sampling Distributions. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 205–213, 2015.

[12] Y. Drori and O. Shamir. The complexity of finding stationary points with stochastic gradient descent. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 2658–2667. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/drori20a.html.

[13] C. Fang, C. J. Li, Z. Lin, and T. Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/file/1543843a4723ed2ab08e18053ae6dc5b-Paper.pdf.

[14] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM J. Optim.*, 23(4):2341–2368, 2013.

[15] R. Gower, N. Loizou, X. Qian, A. Sailanbayev, E. Shulgin, and P. Richtárik. SGD: General analysis and improved rates. In *International Conference on Machine Learning (ICML)*. PMLR, 2019.

[16] T. H. Gronwall. Note on the derivatives with respect to a parameter of the solutions of a system of differential equations. *Ann. of Math. (2)*, 20(4):292–296, 1919. ISSN 0003-486X. doi: 10.2307/1967124. URL https://doi.org/10.2307/1967124.

[17] P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Parallelizing Stochastic Gradient Descent for Least Squares Regression: Mini-batching, Averaging, and Model Misspecification. *Journal of Machine Learning Research*, 18(223):1–42, 2018.

[18] P. Jain, S. Kakade, R. Kidambi, P. Netrapalli, and A. Sidford. Accelerating Stochastic Gradient Descent for Least Squares Regression. In *Proceedings of the 31st Conference On Learning Theory (COLT)*, volume 75, pages 545–604, 2018.

[19] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, 2013.

[20] Y. LeCun, C. Cortes, and C. Burges. "mnist" handwritten digit database, 2010. URL http://yann.lecun.com/exdb/mnist.

[21] K. Lee, A.N. Cheng, E. Paquette, and C. Paquette. Trajectory of Mini-batch Momentum: Batch Size Saturation and Convergence in High-dimensions. *arXiv preprint arXiv:2206.01029*, 2022.

[22] S. Mei and A. Montanari. The generalization error of random features regression: precise asymptotics and the double descent curve. *Comm. Pure Appl. Math.*, 75(4):667–766, 2022. doi: 10.1002/cpa.22008. URL https://doi.org/10.1002/cpa.22008.

[23] E. Moulines and F. Bach. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 24, 2011.

[24] P. Nakkiran, B. Neyshabur, and H. Sedghi. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers. In *International Conference on Learning Representations (ICLR)*, 2021.

[25] D. Needell, N. Srebro, and R. Ward. Stochastic gradient descent, weighted sampling, and the randomized Kaczmarz algorithm. *Math. Program.*, 155(1-2, Ser. A):549–573, 2016. doi: 10.1007/s10107-015-0864-7. URL https://doi.org/10.1007/s10107-015-0864-7.

[26] A. S. Nemirovsky and D. B. and Yudin. *Problem complexity and method efficiency in optimization*. Wiley-Interscience Series in Discrete Mathematics. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson.

[27] Y. Nesterov. *Introductory lectures on convex optimization*. Springer, 2004.

[28] C. Paquette, K. Lee, F. Pedregosa, and E. Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory (COLT)*, volume 134 of *Proceedings of Machine Learning Research*, pages 3548–3626, 2021.

[29] E. Paquette, C. Paquette, B. Adlam, and J. Pennington. Homogenization of SGD in high-dimensions: exact dynamics and generalization properties. *arXiv preprint arXiv:2205.07069*, 2022.

[30] B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 04, 1964.

[31] H. Robbins and S. Monro. A Stochastic Approximation Method. *Ann. Math. Statist.*, 1951. doi: 10.1214/aoms/1177729586. URL https://doi.org/10.1214/aoms/1177729586.

[32] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 2017.

2. Appendix

2.1. Deferred Definitions

Definition 6 *A function $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is quadratic if it is a degree-2 polynomial or equivalently if it can be represented by*

$$\mathcal{R}(x) = \frac{1}{2}\boldsymbol{x}^T \boldsymbol{S}\boldsymbol{x} + \boldsymbol{h}^T\boldsymbol{x} + u, \tag{12}$$

for some $d \times d$ matrix $\boldsymbol{S}$, vector $\boldsymbol{h} \in \mathbb{R}^d$, and scalar $u \in \mathbb{R}$. We assume without loss of generality that the matrix $\boldsymbol{S}$ is symmetric. For any quadratic, define the H^2-norm:

$$||\mathcal{R}||_{H^2} \stackrel{def}{=} |\mathcal{R}(0)| + ||\nabla\mathcal{R}(0)|| + ||\nabla^2\mathcal{R}|| = \|\boldsymbol{S}\| + \|\boldsymbol{h}\| + |u|.$$

We note that under Assumption 1 the empirical risk, f, will have bounded H^2-norm. To execute our exact dynamic of statistics, we require Assumption 5 on the quadratic in the same spirit as Assumption 3.

2.2. Deferred Assumptions

Assumption 2 (Polynomially related) *There is an $\rho \in (0, 1)$ so that*

$$d^\rho \leq n \leq d^{1/\rho}.$$

Assumption 3 *Suppose Ω is a bounded contour in the complex plane enclosing $[0, 1 + \|\boldsymbol{A}\|^2]$ at distance $1/2$. Suppose there is an $\alpha_0 \in (0, \frac{1}{4})$ for which*

1. $\max\limits_{z\in\Omega} \max\limits_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{b}| \leq n^{\alpha_0 - 1/2}$.
2. $\max\limits_{z\in\Omega} \max\limits_{1\leq i\neq j\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_j^T| \leq n^{\alpha_0 - 1/2}$.
3. $\max\limits_{z\in\Omega} \max\limits_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i - \frac{1}{n}\operatorname{tr} R(z; \boldsymbol{A}\boldsymbol{A}^T)| \leq n^{\alpha_0 - 1/2}$.

We use the notation $|\Omega| = \max_{z\in\Omega} |\Omega(z)|$ which we assume is bounded independent of n and d.

Figure 2.2 illustrates that both CIFAR-5M and MNIST (with its 99.99 percentile outliers removed) satisfy Assumption 3.

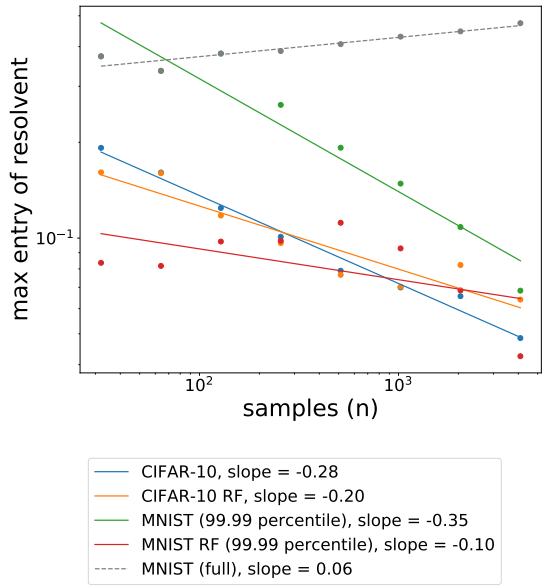

Figure 2: **Maximum off-diagonal entry of the resolvent for CIFAR-5M [24] and MNIST [20] data sets** with features d fixed (3072 and 784, respectively), varying samples $n = 2^k$ for $k = 5, 6, \ldots, 12$. Random features (RF) model was employed with $n_0 = 2000$. In the MNIST (99.99 percentile) data set large resolvent outliers were removed; when outliers not removed, MNIST data set does not satisfy the off-diagonal resolvent condition (Assumption 3 (ii)). For the other data sets, the off-diagonal resolvent condition is satisfied.

Assumption 4 *Let Ω be the same contour as in Assumption 3 and let $\alpha_0 \in (0, \frac{1}{4})$. Then*

$$\max_{z\in\Omega} \max_{1\le i\le d} |\boldsymbol{e}_i^T R(z; \boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_0| \le n^{\alpha_0 - 1/2}.$$

Assumption 5 (Quadratic statistics) *Suppose $\mathcal{R} : \mathbb{R}^d \to \mathbb{R}$ is quadratic, i.e., there is a symmetric matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$, a vector $\boldsymbol{h} \in \mathbb{R}^d$, and a constant $u \in \mathbb{R}$ so that*

$$\mathcal{R}(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{S}\boldsymbol{x} + \boldsymbol{h}^T\boldsymbol{x} + u. \tag{13}$$

We assume that $\mathcal{R}$ satisfies $\|\mathcal{R}\|_{H^2} \le C$ for some C independent of n and d. Moreover we assume the following (for the same Ω and α_0) as in Assumption 3:

$$\max_{z,y\in\Omega} \max_{1\le i\le n} |\boldsymbol{e}_i^T \boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^T\boldsymbol{e}_i - \tfrac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{T}\boldsymbol{A}^T\right)| \le \|\boldsymbol{S}\| n^{\alpha_0 - 1/2} \quad \textit{where} \quad \begin{cases} \boldsymbol{T} \overset{def}{=} R(z)\cdot\boldsymbol{S}\cdot R(y) \\ R(z) \overset{def}{=} R(z; \boldsymbol{A}^T\boldsymbol{A}) \end{cases} \tag{14}$$

2.3. Concentration:In-Distribution Generalization Error.

The following figure illustrates capturing the in-distribution generalization error.

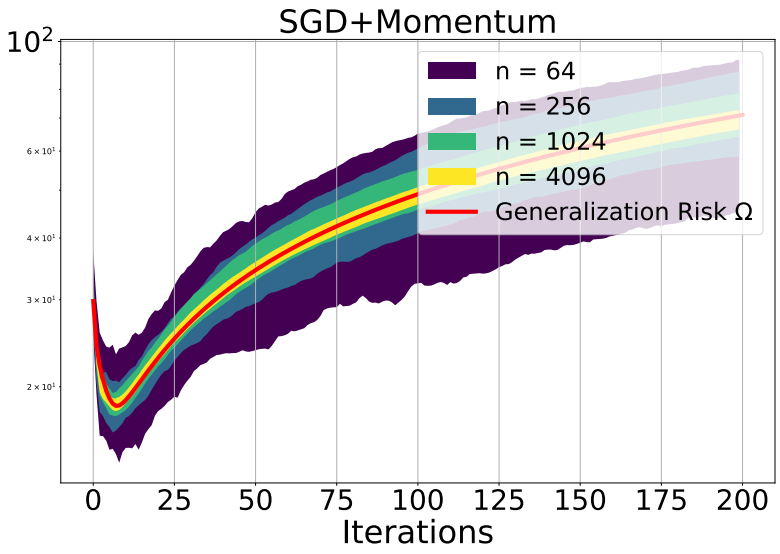

Figure 3: **Concentration of In-Distribution Generalization Error on Gaussian Data.** The *in-distribution* population risk concentrates around $\Omega(t)$ given by Theorem 5. We follow the setup described in Section 1.2.2 with Gaussian data and with mini-batch gradient descent with momentum [30].

2.4. Main Theorem and Proof

We now state the formal version of Theorem 3.

Theorem 7 ((Formal) Dynamics of Mini-Batch, Gradient-Based Methods) *Suppose Assumptions 1, 2, 3, and 4 hold on the data matrix $\boldsymbol{A}$, targets $\boldsymbol{b}$, and initialization $\boldsymbol{x}_0$ with $\alpha_0 \in (0, 1/4)$. Let $\mathcal{R}(\cdot)$ be a quadratic statistic satisfying Assumption 5. Consider the iterates $\{\boldsymbol{x}_t\}_{t=0}^{\infty}$ generated by a mini-batch, gradient-based algorithm (4) with batch size satisfying $\beta/n = \zeta$ for some $\zeta > 0$. For $T > 0$, there exists $\tilde{C} > 0$ such that for any $c > 0$, there exists $D > 0$ satisfying*

$$\Pr\left[\sup_{0\le t\le T}\left\|\begin{pmatrix}\mathcal{L}(\boldsymbol{x}_t)\\ \mathcal{R}(\boldsymbol{x}_t)\end{pmatrix} - \begin{pmatrix}\Psi(t)\\ \Omega(t)\end{pmatrix}\right\| > C(T)n^{-\tilde{C}}\right] \le Dn^{-c}, \tag{15}$$

where $C(T)$ is a constant depending on T, independent of n and d. The functions Ψ and Ω are given as

$$\begin{aligned}\Psi(t) &= \mathcal{L}(\boldsymbol{y}_t) + \zeta(1-\zeta)\sum_{k=1}^{t}\mathcal{K}(t,k;\nabla^2\mathcal{L})\Psi(k-1)\\ \Omega(t) &= \mathcal{R}(\boldsymbol{y}_t) + \zeta(1-\zeta)\sum_{k=1}^{t}\mathcal{K}(t,k;\nabla^2\mathcal{R})\Psi(k-1)\end{aligned} \tag{16}$$

with $\{\boldsymbol{y}_t\}$ *given in* (6) *and where we define the quantities for any* $d \times d$ *matrix* $\boldsymbol{M}$

$$\mathcal{K}(t,k;\boldsymbol{M}) = \frac{1}{n}\operatorname{tr}\left(N_{t,k}(\boldsymbol{H})\boldsymbol{M}N_{t,k}(\boldsymbol{H})\boldsymbol{A}^T\boldsymbol{A}\right), \qquad t \geq 0 \qquad \textit{and} \qquad 0 \leq k \leq t \tag{17}$$

and where the sequence of polynomials $\{N_{k,j}\}_{k\in[0,\infty),j\leq k}$ *are defined recursively by*

$$N_{0,0}(\boldsymbol{H}) = \boldsymbol{0} \quad \textit{and} \quad N_{k+1,j}(\boldsymbol{H}) = \begin{cases} \sum_{i=j}^{k} \zeta c_{k,i}\boldsymbol{H}N_{i,j}(\boldsymbol{H}) - c_{k,j-1}, & \textit{if } j = 1,\dots,k+1 \\ \sum_{i=0}^{k} \zeta c_{k,i}\boldsymbol{H}N_{i,0}(\boldsymbol{H}), & \textit{if } j = 0 \\ -c_{k,k}, & \textit{if } j = k+1 \end{cases},$$

where $c_{k,j}$ *are the mini-batch coefficients and* $\boldsymbol{H} \overset{\textit{def}}{=} \delta\boldsymbol{I} + \boldsymbol{A}^\top\boldsymbol{A}$.

The proof of Theorem 7 consists of constructing two deterministic quantities Ψ_t and Ω_t such that the difference with $\mathcal{L}(\boldsymbol{x}_t)$ and $\mathcal{R}(\boldsymbol{x}_t)$, respectively, can be expressed as a summation of error terms that vanish in the high-dimensional limit. The proofs that these terms are small will be deferred to the next section.

To begin, recall the iterates of any mini-batch gradient-based method under any quadratic function $\mathcal{R}(\cdot)$ (Assumption 5), are expressed as

$$\mathcal{R}(\boldsymbol{x}_k) = \frac{1}{2}\boldsymbol{x}_k^T\boldsymbol{S}\boldsymbol{x}_k + \boldsymbol{h}^T\boldsymbol{x}_k + \boldsymbol{u}, \tag{18}$$

where the matrix $\boldsymbol{S} \in \mathbb{R}^{d\times d}$ is symmetric, $\boldsymbol{h} \in \mathbb{R}^d$ is a vector, and $u \in \mathbb{R}$ is a constant. Using some simplifications, the iterates can also be expressed by

$$\boldsymbol{x}_k = \boldsymbol{y}_k + \sum_{j=0}^{k} N_{k,j}(\boldsymbol{H})\mathring{\boldsymbol{M}}_j \tag{19}$$

where the iterates $\boldsymbol{y}_k$ satisfy $\boldsymbol{y}_{k+1} = \boldsymbol{x}_0 + \sum_{j=0}^{k}\zeta c_{kj}\nabla f(\boldsymbol{y}_k)$ and where we define $\mathring{\boldsymbol{M}}_j$

$$\begin{aligned} \mathring{\boldsymbol{M}}_j &\overset{\text{def}}{=} \mathbb{E}\left[\boldsymbol{A}^T\boldsymbol{P}_j\left(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}\right)\middle|\mathcal{F}_{j-1}\right] - \boldsymbol{A}^T\boldsymbol{P}_j\left(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}\right) \\ &= \zeta\boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}) - \boldsymbol{A}^T\boldsymbol{P}_j(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}) \\ &\text{where } \boldsymbol{P}_{j+1} \overset{\text{def}}{=} \sum_{i\in B_j}\boldsymbol{e}_i\boldsymbol{e}_i^\top \end{aligned} \tag{20}$$

to be the *martingale increment*. The martingale incremrements represent a Doob's decomposition that allows us to handle the stochasticity generated from the mini-batches.

Since everything is evaluated at $\boldsymbol{H} = \delta\mathbf{I} + \boldsymbol{A}^T\boldsymbol{A}$, we will often suppress the $\boldsymbol{H}$ and write $\mathbf{N}_{k,j} \overset{\text{def}}{=} N_{k,j}(\boldsymbol{H})$. Moreover for any matrix $\boldsymbol{M}$, we introduce $\tilde{\mathbf{N}}_{t,k}(\boldsymbol{M};\boldsymbol{H}) \overset{\text{def}}{=} N_{t,k}(\boldsymbol{H})\boldsymbol{M}N_{t,k}(\boldsymbol{H})$. Now using (19), we get an expression for the quadratic statistic $\mathcal{R}$ as follows:

$$\begin{aligned}\mathcal{R}(\boldsymbol{x}_t) &= \mathcal{R}(\boldsymbol{y}_t) + \nabla\mathcal{R}(\boldsymbol{y}_t)^T \left(\sum_{k=1}^{t} \mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k\right) + \frac{1}{2}\left(\sum_{k=1}^{t} \mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k\right)^T (\nabla^2\mathcal{R}) \left(\sum_{k=1}^{t} \mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k\right) \\ &= \mathcal{R}(\boldsymbol{y}_t) + \nabla\mathcal{R}(\boldsymbol{y}_t)^T \left(\sum_{k=1}^{t} \mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k\right) + \frac{1}{2}\sum_{k=1}^{t} \mathring{\boldsymbol{M}}_k^T \mathbf{N}_{t,k}(\nabla^2\mathcal{R})\mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k \\ &\quad + \sum_{k_1>k_2} \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{N}_{t,k_1}(\nabla^2\mathcal{R})\mathbf{N}_{t,k_2}\mathring{\boldsymbol{M}}_{k_2}.\end{aligned} \tag{21}$$

Here we decomposed the Hessian term, $\nabla^2\mathcal{R}$, into two terms. The first term is the "on diagonal" term, that is, when $\mathring{\boldsymbol{M}}_j$ and $\mathring{\boldsymbol{M}}_k$ have the same index in the two summations. When $\mathring{\boldsymbol{M}}_k$ and $\mathring{\boldsymbol{M}}_j$ have different indices, we put them into the off-diagonal term. Next we introduce notation to simplify the computations. In particular, we define

$$\boldsymbol{w}_t \overset{\text{def}}{=} \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b} \quad \text{and} \quad \boldsymbol{w}_{t,i} = (\boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b})_i \quad \text{for } i = 1, \dots, n,$$

to be an entry of the empirical loss function.

We are now going to identify the leading order behavior of $\mathcal{R}(\boldsymbol{x}_t)$. For this, we will see that the gradient term ($\nabla\mathcal{R}$) and the off-diagonal term are of lower order in comparison to $\mathcal{R}(\boldsymbol{y}_t)$. They will be shown to vanish as $d \to \infty$. As such we denote these two terms as error terms,

$$\mathcal{E}_t^{\nabla}(\mathcal{R}) \overset{\text{def}}{=} \nabla\mathcal{R}(\boldsymbol{y}_t)^T \left(\sum_{k=1}^{t} \mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k\right) \tag{22}$$

$$\mathcal{E}_t^{\nabla^2\text{-Off}}(\mathcal{R}) \overset{\text{def}}{=} \sum_{k_1>k_2}^{t} \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{N}_{t,k_1}(\nabla^2\mathcal{R})\mathbf{N}_{t,k_2}\mathring{\boldsymbol{M}}_{k_2}. \tag{23}$$

The exact statement and subsequent proof that these terms are small can be found in Proposition 2.1 and Proposition 2.3, respectively.

For the "on-diagonal" Hessian term, there is a component of it which survives the limit in n. To isolate this component, we must expand the on-diagonal term using the definition of $\mathring{\boldsymbol{M}}$. We defer this computation until Section **??** as it is quite involved. As such, we write this term as an error term plus its leading behavior:

$$\begin{aligned}\frac{1}{2}\sum_{k=1}^{t} \mathring{\boldsymbol{M}}_k^T \mathbf{N}_{t,k}(\nabla^2\mathcal{R})\mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k = \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n} \frac{(\zeta - \zeta^2)}{n} \operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{t,k}(\nabla^2 R; \boldsymbol{H})\boldsymbol{A}^T\right) \boldsymbol{w}_{k-1,\ell}^2 \\ + \mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R}).\end{aligned} \tag{24}$$

Here the error term is given by

$$\begin{aligned}\mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R}) \overset{\text{def}}{=} \frac{1}{2}\sum_{k=1}^{t} \mathring{\boldsymbol{M}}_k^T \mathbf{N}_{t,k}(\nabla^2\mathcal{R})\mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k \\ - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n} \frac{(\zeta - \zeta^2)}{n} \operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{(t,k)}(\nabla^2 R; \boldsymbol{H})\boldsymbol{A}^T\right) \boldsymbol{w}_{k-1,\ell}^2\end{aligned} \tag{25}$$

The statement and proof that $\mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R})$ vanishes in d is defered to Proposition 2.2.

Using (26), we write the risk as

$$\begin{aligned}\mathcal{R}(\boldsymbol{x}_t) = \mathcal{R}(\boldsymbol{y}_t) + \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{t,k}(\nabla^2 R;\boldsymbol{H})\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2 \\ + \mathcal{E}_t^{\nabla}(\mathcal{R}) + \mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R}) + \mathcal{E}_t^{\nabla^2-\text{Off}}(\mathcal{R}).\end{aligned} \tag{26}$$

The terms, $\mathcal{E}_t^{\nabla}(\mathcal{R})$, $\mathcal{E}_t^{\nabla^2\text{-Diag}}(\mathcal{R})$, and $\mathcal{E}_t^{\nabla^2-\text{Off}}(\mathcal{R})$, are error terms that vanish when n or d is sufficiently large. We often will drop the $\mathcal{R}$ in the definition of the error terms when it is clear.

In order to prove Theorem 7, we require a discrete version of Gronwall's inequality, a proof of which can be found in [16].

Lemma 8 (Discrete Gronwall Inequality) *Let $\mathcal{E}(k)$ and $\tilde{\mathcal{K}}(k)$ be non-negative, non-decreasing sequences and $F(k)$ be a non-negative sequence defined for $k = 0, 1, 2, \ldots$. For any $T > 0$, suppose the sequences satisfy*

$$F(T) \leq \mathcal{E}(T) + \tilde{\mathcal{K}}(T)\sum_{0\leq k<T} F(k), \tag{27}$$

and $F(0) \leq \mathcal{E}(0)$. Then for any $T \geq 0$,

$$F(T) \leq \mathcal{E}(T) + T\cdot\mathcal{E}(T)\tilde{\mathcal{K}}(T)\exp\left(T\cdot\tilde{\mathcal{K}}(T)\right). \tag{28}$$

Now we are ready to prove our main result, Theorem 7, for the dynamics of any mini-batch gradient-based method. Recall the vector $\boldsymbol{w}_t = \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b}$ and empirical loss function

$$\mathcal{L}(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2.$$

It will be convenient to work with a stopped process based on the stopping time ϑ:

$$\vartheta \stackrel{\text{def}}{=} \inf\{t \geq 0 : ||\boldsymbol{w}_t|| > n^\theta\} \tag{29}$$

for some $\theta > 0$ to be determined. We define the stopped process of the iterates and the residual vector to be

$$\boldsymbol{x}_t^\vartheta \stackrel{\text{def}}{=} \boldsymbol{x}_{t\wedge\vartheta} \quad \text{and} \quad \boldsymbol{w}_t^\vartheta \stackrel{\text{def}}{=} \boldsymbol{w}_{t\wedge\vartheta} \quad t \geq 0.$$

We will first prove that Theorem 7 holds for the stopped process $\boldsymbol{x}_t^\vartheta$. Then we will remove the stopping time ϑ and get the result for the entire sequence $\boldsymbol{x}_t$.

Proof [Theorem 7] Let $\{N_{t,k}\}$ be the noise polynomial that corresponds to the mini-batch gradient-based iterates $\boldsymbol{x}_t$ (see Definition 2). Throughout the proof, we will suppress the $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{A}^T + \delta\mathbf{I}$, use $\mathbf{N}_{t,k} = N_{t,k}(\boldsymbol{H})$, and let $\tilde{\mathbf{N}}_{t,k}(\boldsymbol{M};\boldsymbol{H}) = N_{t,k}(\boldsymbol{H})\boldsymbol{M}N_{t,k}(\boldsymbol{H})$ for all $t \geq 0$.

First we show that the stopped process $\boldsymbol{x}_t^\vartheta$ under the empirical loss is close to the stopped process under $\Psi(t)$, that is, $\mathcal{L}(\boldsymbol{x}_t^\vartheta)$ is close to $\Psi(t \wedge \vartheta)$. For this, we apply (26) replacing $\mathcal{R}$ with $\mathcal{L}$ and set $t \mapsto (t \wedge \vartheta)$, so that

$$\begin{aligned}\mathcal{L}(\boldsymbol{x}_t^\vartheta) = \mathcal{L}(\boldsymbol{y}_t^\vartheta) + \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{t\wedge\vartheta,k}(\nabla^2\mathcal{L};\boldsymbol{H})\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2 \\ + \mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{L}) + \mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{L}) + \mathcal{E}_{t\wedge\vartheta}^{\nabla^2-\text{Off}}(\mathcal{L}).\end{aligned} \tag{30}$$

From now on we suppress the argument of $\tilde{\mathbf{N}}_{t,k}$, that is, we write $\tilde{\mathbf{N}}_{t,k} = \tilde{\mathbf{N}}_{t,k}(\nabla^2\mathcal{L};\boldsymbol{H})$.

Comparing the two processes, we have

$$\begin{aligned}\left|\mathcal{L}(\boldsymbol{x}_t^\vartheta) - \Psi(t\wedge\vartheta)\right| &\leq \left|\sum_{k=1}^{t\wedge\vartheta}(\zeta-\zeta^2)\mathcal{K}\left(t\wedge\vartheta,k;\nabla^2\mathcal{L}\right)\Psi(k-1)\right. \\ &\qquad \left. - \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\right)(\boldsymbol{w}_{k-1,\ell}^\vartheta)^2\right| \\ &\quad + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{L})\right| \\ &\leq (\zeta-\zeta^2)\sum_{k=1}^{t\wedge\vartheta}\left|\mathcal{K}(t\wedge\vartheta,k;\nabla^2\mathcal{L})\right|\cdot\left|\mathcal{L}(\boldsymbol{x}_{k-1}^\vartheta) - \Psi\left((k-1)\wedge\vartheta\right)\right| \\ &\quad + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{L})\right|\end{aligned} \tag{31}$$

where in the last inequality, we used the fact that $\mathcal{L}(\boldsymbol{x}_k^\vartheta) = \frac{1}{2}\|\boldsymbol{w}_k^\vartheta\|^2$ and the kernel $\mathcal{K}(t\wedge\vartheta,k;\boldsymbol{M}) = \frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{t\wedge\vartheta,k}(\boldsymbol{M};\boldsymbol{H})\boldsymbol{A}^T\right)$. In particular, we obtain

$$\begin{aligned}&\max_{0\leq t\leq T}\left|\mathcal{L}(\boldsymbol{x}_t^\vartheta) - \Psi(t\wedge\vartheta)\right| \\ &\quad\leq \sum_{k=1}^{T}\left((\zeta-\zeta^2)\max_{0\leq s\leq T,0\leq k\leq s}\left|\mathcal{K}(s\wedge\vartheta,k\wedge\vartheta;\nabla^2\mathcal{L})\right|\right)\left(\max_{0\leq s\leq k}\left|\mathcal{L}(\boldsymbol{x}_s^\vartheta) - \Psi(s\wedge\vartheta)\right|\right) \\ &\quad + \max_{0\leq t\leq T}\left(\left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{L})\right|\right).\end{aligned} \tag{32}$$

We define the following terms in order to rewrite (32) in a recursive form

$$F(T) \overset{\text{def}}{=} \max_{0\leq t\leq T}\left|\mathcal{L}(\boldsymbol{x}_t^\vartheta) - \Psi(t\wedge\vartheta)\right|, \quad \mathcal{E}(T;\mathcal{L}) \overset{\text{def}}{=} \max_{0\leq t\leq T}\left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{L})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{L})\right|,$$
$$\text{and} \quad \tilde{\mathcal{K}}(T;\mathcal{L}) \overset{\text{def}}{=} (\zeta-\zeta^2)\max_{0\leq s\leq T,0\leq k\leq s}\left|\mathcal{K}(s\wedge\vartheta,k\wedge\vartheta;\nabla^2\mathcal{L})\right|.$$

Rewriting (32) in terms of these quantities gives the recursive form that resembles Gronwall inequality:

$$F(T) \leq \mathcal{E}(T;\mathcal{L}) + \sum_{0\leq k<T} \tilde{\mathcal{K}}(T;\mathcal{L})F(k) \qquad \text{for } T \geq 0. \tag{33}$$

Note that $\{F(T) : T \in \mathbb{N}\}, \{\mathcal{E}(T;\mathcal{L}) : T \in \mathbb{N}\}$, and $\{\tilde{\mathcal{K}}(T;\mathcal{L}) : T \in \mathbb{N}\}$ are non-negative sequences. Moreover $\mathcal{E}(T;\mathcal{L})$ and $\tilde{\mathcal{K}}(T;\mathcal{L})$ are non-decreasing in T, and

$$F(0) = |\mathcal{L}(\boldsymbol{x}_0) - \Psi(0)| = 0 \leq \mathcal{E}(0;\mathcal{L}).$$

Thus, the assumptions of Lemma 8 hold and we have that

$$\begin{aligned} F(T) &\leq \mathcal{E}(T;\mathcal{L}) + T\cdot\mathcal{E}(T;\mathcal{L})\tilde{\mathcal{K}}(T)\exp\left(T\cdot\tilde{\mathcal{K}}(T;\mathcal{L})\right) \\ &\leq C(T;\mathcal{L})\mathcal{E}(T;\mathcal{L}), \end{aligned} \tag{34}$$

where $C(T;\mathcal{L})$ is a constant independent of n and d. Now we show that the term $\mathcal{E}(T;\mathcal{L})$ is small as $n, d \to \infty$. This is done by applying Proposition 2.1, Proposition 2.2, and Proposition 2.3. Given some $\alpha_0 \in (0, 1/4)$, we assign values of $\alpha, \alpha', \theta, \eta$, and $\hat{\delta}$ so that each of the terms in $\mathcal{E}(T;\mathcal{L})$ becomes vanishingly small as n, d tend to infinity. Specifically, we assign the following values

$$\begin{aligned} &\alpha \overset{\text{def}}{=} \frac{\alpha_0}{2} + \frac{1}{8}, \qquad \alpha' \overset{\text{def}}{=} \frac{\alpha_0}{4} + \frac{3}{16}, \qquad \hat{\delta} \overset{\text{def}}{=} \frac{\alpha_0}{8} + \frac{7}{32}, \\ &\theta \overset{\text{def}}{=} \frac{1}{4}\left(-\frac{\alpha_0}{4} + \frac{1}{16}\right), \quad \text{and} \qquad \eta \overset{\text{def}}{=} \frac{\hat{\delta} - \alpha}{2}. \end{aligned} \tag{35}$$

Using these values for Proposition 2.1, Proposition 2.2, and Proposition 2.3 yields

$$\begin{aligned} \mathcal{E}(T;\mathcal{L}) &\overset{\text{def}}{=} \max_{0\leq t\leq T} \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}\right| + \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}\right| + \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}\right| \\ &\leq \max_{0\leq t\leq T} \left|\mathcal{E}^{\nabla}_{t\wedge\vartheta}\right| + \max_{0\leq t\leq T} \left|\mathcal{E}^{\nabla^2\text{-Diag}}_{t\wedge\vartheta}\right| + \max_{0\leq t\leq T} \left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}\right| \\ &\leq C(T;\mathcal{L})n^{-c} \qquad \text{w.o.p.} \end{aligned} \tag{36}$$

for some $c > 0$. Hence, we have the first result (for the stopped process), that is, with overwhelming probability

$$\max_{0\leq t\leq T} \left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \leq C(T;\mathcal{L})n^{-c}, \tag{37}$$

for some constant $c > 0$ and $C(T;\mathcal{L})$ which is independent of n and d.

We repeat the argument for the statistic $\mathcal{R}$. Using (26), with the stopped process $t \mapsto t \wedge \vartheta$ and the definition of Ω (15) on the stopped process, we observe

$$\begin{aligned}\left|\Omega(t\wedge\vartheta) - \mathcal{R}(\boldsymbol{x}_t^\vartheta)\right| \leq & \left| \sum_{k=1}^{t\wedge\vartheta} (\zeta - \zeta^2)\mathcal{K}\left((t\wedge\vartheta) - k; \nabla^2\mathcal{R}\right)\Psi(k-1) \right. \\ & \left. - \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n} \frac{(\zeta-\zeta^2)}{n} \operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\right)(\boldsymbol{w}_{k-1,\ell}^\vartheta)^2 \right| \\ & + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{R})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{R})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{R})\right| \\ \leq & (\zeta-\zeta^2)\sum_{k=1}^{t\wedge\vartheta}\left|\frac{1}{n}\operatorname{tr}\left(\tilde{\mathbf{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right| \cdot \left|\Psi((k-1)\wedge\vartheta) - \mathcal{L}(\boldsymbol{x}_{k-1}^\vartheta)\right| \\ & + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{R})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{R})\right| + \left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{R})\right|\end{aligned} \tag{38}$$

where we used the fact that $\mathcal{L}(\boldsymbol{x}_k^\vartheta) = \frac{1}{2}\|\boldsymbol{w}_k^\vartheta\|^2$ and $\tilde{\mathbf{N}}_{t,k} \overset{\text{def}}{=} \tilde{\mathbf{N}}_{t,k}(\nabla^2\mathcal{R}; \boldsymbol{H})$. Taking the supremum on both sides yields,

$$\begin{aligned}\max_{0\leq t\leq T}\left|\Omega(t\wedge\vartheta) - \mathcal{R}(\boldsymbol{x}_t^\vartheta)\right| \leq & \max_{0\leq t\leq T}\left|\Psi(t\wedge\vartheta) - \mathcal{L}(\boldsymbol{x}_t^\vartheta)\right| \sum_{k=1}^{t\wedge\vartheta}\left|\frac{1}{n}\operatorname{tr}\left(\tilde{\mathbf{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right| \\ & + \max_{0\leq t\leq T}\left|\mathcal{E}_{t\wedge\vartheta}^{\nabla}(\mathcal{R})\right| + \max_{0\leq t\leq T}\left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Diag}}(\mathcal{R})\right| + \max_{0\leq t\leq T}\left|\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}(\mathcal{R})\right| \\ \leq & C(T;\mathcal{L})n^{-c}\max_{0\leq t\leq T}\left\{\sum_{k=1}^{t\wedge\vartheta}\left|\frac{1}{n}\operatorname{tr}\left(\tilde{\mathbf{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)\right|\right\} + \mathcal{E}(T;\mathcal{R}).\end{aligned} \tag{39}$$

Here we used (37) with some $c > 0$. By the assumptions on the data matrix $\boldsymbol{A}$, we have that $\frac{1}{n}\operatorname{tr}\left(\tilde{\mathbf{N}}_{t\wedge\vartheta,k}\boldsymbol{A}^T\boldsymbol{A}\right)$ is bounded independent of n. Therefore, we just need to show that the error term $\mathcal{E}(T;\mathcal{R})$ vanishes as $n, d \to \infty$. This follows exactly the same proof as $\mathcal{E}(T;\mathcal{L})$. Using the same values for $\alpha, \alpha', \theta, \eta$, and $\hat{\delta}$ as in (35) and replacing $\mathcal{L}$ with $\mathcal{R}$ in Proposition 2.1, Proposition 2.2, and Proposition 2.3, we have that

$$\mathcal{E}(T;\mathcal{R}) \leq C(T,\mathcal{R})n^{-c} \quad \text{w.o.p}$$

for some $c > 0$ and constant $C(T;\mathcal{R})$ independent of n and d.

Hence we have shown, with overwhelming probability,

$$\max_{0\leq t\leq T}\left|\Omega(t\wedge\vartheta) - \mathcal{R}(\boldsymbol{x}_t^\vartheta)\right| \leq C(T;\mathcal{L};\mathcal{R})n^{-c} \tag{40}$$

for some constant $c > 0$ and constant $C(T;\mathcal{L};\mathcal{R})$ which is independent of n and d and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$.

To finish off the proof, we now show that the stopping time satisfies $\vartheta > T$ with overwhelming probability. For sufficiently large n, by the definition of the stopping time, the fact that $\mathcal{L}(\boldsymbol{x}_t) =$

$\frac{1}{2}\|\boldsymbol{w}_t\|_2^2$, and using the constant $c > 0$ in (37) yields

$$\begin{aligned}\Pr(\vartheta > T) &\geq \Pr\left(\max_{0\leq t\leq T}\left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \leq \frac{1}{2}n^{2\theta} - \max_{0\leq t\leq T}\Psi(t)\right)\\ &\geq 1 - \Pr\left(\max_{0\leq t\leq T}\left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \geq \frac{1}{2}n^{2\theta} - \max_{0\leq t\leq T}\Psi(t)\right)\\ &\geq 1 - \Pr\left(\max_{0\leq t\leq T}\left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| \geq n^{-c}\right)\\ &\geq 1 - Dn^{-\tilde{c}}\end{aligned} \tag{41}$$

for some constant $\tilde{c} > 0$ where the last line follows from the result of (37). We note that $\max_{0\leq t\leq T}\Psi(t)$ is independent of n and d and the maximum is taken over a finite set so the maximum is finite. Using (37) and (41), we get

$$\begin{aligned}&\Pr\left(\max_{0\leq t\leq T}|\mathcal{L}(\boldsymbol{x}_t) - \Psi(t)| > n^{-c}\right) = \Pr\left(\max_{0\leq t\leq T}|\mathcal{L}(\boldsymbol{x}_t) - \Psi(t)|\left(\mathbf{1}_{\{\vartheta>T\}} + \mathbf{1}_{\{\vartheta\leq T\}}\right) > n^{-c}\right)\\ &\leq \Pr\left(\max_{0\leq t\leq T}|\mathcal{L}(\boldsymbol{x}_t) - \Psi(t)|\,\mathbf{1}_{\{\vartheta\leq T\}} > n^{-c}\right) + \Pr\left(\max_{0\leq t\leq T}\left|\mathcal{L}(\boldsymbol{x}_t^{\vartheta}) - \Psi(t\wedge\vartheta)\right| > n^{-c}\right)\\ &\leq Dn^{-\tilde{c}}\end{aligned} \tag{42}$$

for some constants D and $c > 0$ that are independent of n and d. This shows

$$\max_{0\leq t\leq T}\ |\Psi(t) - \mathcal{L}(\boldsymbol{x}_t)| \leq n^{-c}$$

holds with overwhelming probability. Similarly, by replacing $\mathcal{L}$ with $\mathcal{R}$, we can show

$$\max_{0\leq t\leq T}\ |\Omega(t) - \mathcal{R}(\boldsymbol{x}_t)| \leq n^{-c}$$

holds with overwhelming probability which gives us what we need. ■

2.4.1. CONTROLLING THE MARTINGALE ERRORS

The martingale errors $\mathcal{E}_t^{\nabla}$, $\mathcal{E}_t^{\nabla^2\text{-Diag}}$, and $\mathcal{E}_t^{\nabla^2\text{-Off}}$, (22), (23), and (25) respectively, arise due to the randomness in the algorithm itself. These errors are small, in part, because the left-singular vector matrix of $\boldsymbol{A}$ is delocalized (Assumption 3). Estimating that the error generated by these martingales requires some substantial build-up (Section 2.4.2). For reference, the three martingale errors are

$$\begin{aligned}\mathcal{E}_t^{\nabla} &= \nabla\mathcal{R}(\boldsymbol{y}_t)^T\left(\sum_{k=1}^{t}\mathbf{N}_{t,k}\mathring{\boldsymbol{M}}_k\right)\\ \mathcal{E}_t^{\nabla^2\text{-Off}} &= \sum_{k_1>k_2}^{t}\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{N}_{t,k_1}(\nabla^2\mathcal{R})\mathbf{N}_{t,k_2}\mathring{\boldsymbol{M}}_{k_2}\\ \mathcal{E}_t^{\nabla^2\text{-Diag}} &= \frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\tilde{\mathbf{N}}_{t,k}\mathring{\boldsymbol{M}}_k - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\tilde{\mathbf{N}}_{(t,k)}\boldsymbol{A}^T\right)\boldsymbol{w}_{k-1,\ell}^2,\end{aligned} \tag{43}$$

where we write $\mathbf{N}_{k,j} \stackrel{\text{def}}{=} N_{k,j}(\boldsymbol{H})$ and $\tilde{\mathbf{N}}_{t,k} \stackrel{\text{def}}{=} N_{t,k}(\boldsymbol{H})(\nabla^2\mathcal{R})N_{t,k}(\boldsymbol{H})$. The incremental martingale $\mathring{\boldsymbol{M}}$ (20) is

$$\mathring{\boldsymbol{M}}_j = \zeta \boldsymbol{A}^T(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}) - \boldsymbol{A}^T\boldsymbol{P}_j(\boldsymbol{A}\boldsymbol{x}_{j-1} - \boldsymbol{b}). \tag{44}$$

where $\boldsymbol{P}_j$ is a random projection matrix.

Throughout this section, we normalize our data matrix $\boldsymbol{A}$ so that it has row sum equals 1, namely

$$||\boldsymbol{A}_i||_2 = 1, \quad i \in [n] \tag{45}$$

without loss of generality. In order to control the fluctuations of these martingales, we will need to make an *a priori* estimate that effectively shows that the iterates remain bounded. Thus, we recall the introduction of a stopping time, for any fix $\theta > 0$, by

$$\vartheta_\theta = \inf\left\{t \geq 0 \,:\, \|\boldsymbol{w}_t\| > n^\theta\right\}, \quad \text{where } \boldsymbol{w}_t = \boldsymbol{A}\boldsymbol{x}_t - \boldsymbol{b}.$$

The choice of θ will be determined later, and as such, we simplify notation by $\vartheta \stackrel{\text{def}}{=} \vartheta_\theta$. It will be convenient to work under the stopped processes, $\boldsymbol{x}_t^\vartheta = \boldsymbol{x}_{t\wedge\vartheta}$ and $\boldsymbol{w}_t^\vartheta = \boldsymbol{w}_{t\wedge\vartheta}$.

In what follows, we will prove three propositions (stated below) showing that each of the martingale error terms in (43) are small starting with $\mathcal{E}_t^\nabla$.

Proposition 2.1 *For all $\alpha' > \alpha_0 + \theta$, the following holds with overwhelming probability*

$$\max_{0\leq t\leq T} |\mathcal{E}^\nabla_{t\wedge\vartheta}| \leq n^{\alpha'-1/2}. \tag{46}$$

In Section 2.4.4, we will show that the on-diagonal error term is small. This will involve using the *key lemma*.

Proposition 2.2 (On-diagonal error term small) *Let $\alpha_0 \in (0, 1/4)$ as specified in Assumption 4. For any α, α', and θ satisfying $0 < \theta < \alpha' - \alpha$ and $\alpha_0 + \theta < \alpha < \alpha'$ we have*

$$\max_{0\leq t\leq T} |\mathcal{E}^{\nabla^2-\text{Diag}}_{t\wedge\vartheta}| \leq C(T)n^{2\alpha'-1/2} \quad \textit{w.o.p.}$$

for some constant $C(T)$ that is independent of n and d.

Finally, in Section 2.5, we show that the off-diagonal error term is small.

Proposition 2.3 *For all α and δ such that $\alpha_0 + \theta < \alpha < \delta < 1/4$ and $0 < \theta < \frac{1}{4} - \delta$ and for any η satisfying $0 < \eta < \delta - \alpha$ we have that the quadratic-off-diagonal term satisfies*

$$\max_{0\leq t\leq T} \left|\mathcal{E}^{\nabla^2\text{-}\mathit{Off}}_{t\wedge\vartheta}\right| = C(T)\cdot n^{\alpha-\delta+\eta} \qquad \textit{w.o.p.,} \tag{47}$$

for some constant $C(T) > 0$ that is independent of n and d and only depends on $||\Omega||$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, γ and T.

2.4.2. GENERAL MARTINGALE RESULTS

Proposition 2.4 (Resolvent and Bounded Entries) *Fix a constant $T > 0$. Suppose Assumptions 3 hold for the closed, bounded contour Ω. For any $\alpha > \alpha_0 + \theta$,*

$$\begin{aligned} \max_{0\le t\le T}\max_{z\in\Omega}\max_{1\le i\le n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^\vartheta| &\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,T)n^{\alpha-1/2} \\ \textit{and}\quad \max_{0\le t\le T}\max_{1\le i\le n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^\vartheta| &\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,T)n^{\alpha-1/2} \end{aligned} \tag{48}$$

with overwhelming probability (conditioned on $\mathcal{F}_0$), where C is constant depending on $\|\Omega\|$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, and time T, but independent of n and d.

We immediately get a consequence which yields a bound on the individual entries of $\boldsymbol{w}_t^\vartheta$.

Lemma 9 (Coordinates are Small in n) *Suppose the assumptions of Proposition 2.4 hold with some $T \ge 0$. For all $\alpha > \alpha_0 + \theta$,*

$$\begin{aligned} \max_{0\le t\le T}\max_{1\le i\le n} |\boldsymbol{e}_i^T\boldsymbol{w}_t^\vartheta| &\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,T)n^{\alpha-1/2} \\ \textit{and}\quad \max_{0\le t\le T}\max_{1\le i\le n} |\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{x}_t^\vartheta| &\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,T)n^{\alpha-1/2} \end{aligned} \tag{49}$$

with overwhelming probability (conditioned on $\mathcal{F}_0$).

Proof [Proof of Lemma 9] From Cauchy's integral formula, we express the i-th entry of $\boldsymbol{A}\boldsymbol{x}_t^\vartheta$ and $\boldsymbol{w}_t^\vartheta$, respectively, as

$$\begin{aligned} |\boldsymbol{e}_i^T\boldsymbol{A}\boldsymbol{x}_t^\vartheta| &= \left|\frac{-1}{2\pi i}\oint_\Omega \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^\vartheta\,\mathrm{d}z\right| \le \frac{|\Omega|}{2\pi}\max_{1\le i\le n}\max_{z\in\Omega}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^\vartheta\right| \\ |\boldsymbol{e}_i^T\boldsymbol{w}_t^\vartheta| &= \left|\frac{-1}{2\pi i}\oint_\Omega \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^\vartheta\,\mathrm{d}z\right| \le \frac{|\Omega|}{2\pi}\max_{1\le i\le n}\max_{z\in\Omega}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^\vartheta\right| \end{aligned} \tag{50}$$

The result then immediately follows after applying Proposition 2.4. ■

Before showing the proof of Proposition 2.4, we state a *Bernstein-type concentration result* for sampling without replacement. This result says the randomness from mini-batch sampling does not deviate too much from the "expectation". It will be used to show Proposition 2.4.

Proposition 2.5 (Bernstein concentration, Proposition 1.4 [5]) *Let $\mathcal{X} = (x_1, \cdots, x_n)$ be a finite population of n points and $X_1, \cdots, X_\beta$ be a random sample drawn without replacement from $\mathcal{X}$. Let*

$$a = \min_{1\le i\le n} x_i \textit{ and } b = \max_{1\le i\le n} x_i.$$

Also let

$$\mu = \frac{1}{n}\sum_{i=1}^n x_i \textit{ and } \sigma^2 = \frac{1}{n}\sum_{i=1}^n (x_i-\mu)^2$$

be the mean and variance of $\mathcal{X}$, respectively. Then for all $\epsilon > 0$,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{i=1}^\beta X_i - \mu \ge \epsilon\right) \le \exp\left(-\frac{\beta\epsilon^2}{2\sigma^2 + (2/3)(b-a)\epsilon}\right).$$

We now prove the general bound on the entries of $\boldsymbol{w}_t^{\vartheta}$.

Proof [Proof of Proposition 2.4] For any fixed $\alpha > \alpha_0 + \theta$, define an increasing sequence $\alpha(t)$ where $t = 0, 1, \ldots, T$ such that $\alpha_0 + \theta < \alpha(0) < \alpha(1) < \ldots < \alpha(T) \leq \alpha$. We will show for any $0 \leq t \leq T$

$$\begin{aligned} \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, t)n^{\alpha(t)-1/2} \\ \text{and} \quad \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, t)n^{\alpha(t)-1/2}, \end{aligned} \tag{51}$$

where C is a constant depending on $|\Omega|$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, and t but it is independent of n and d.

For $t = 0$, a simple computations yield

$$\begin{aligned} \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^{\vartheta}| &= \max_{z\in\Omega}\max_{1\leq i\leq n} \left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{A}\boldsymbol{x}_0 - \boldsymbol{b})\right| \\ &\leq \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{A}\boldsymbol{x}_0| + \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{b}|. \end{aligned}$$

Therefore, for the $\boldsymbol{w}_0^{\vartheta}$ term, the inequality (51) follows after applying Assumption 3 and Assumption 4. By setting $\boldsymbol{b} = 0$ in the inequality above, we get that (51) holds for $\boldsymbol{A}\boldsymbol{x}_0^{\vartheta}$.

For $1 \leq t \leq \vartheta$, we prove by induction. We already showed that

$$\begin{aligned} \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, 0)(n^{\alpha(0)-1/2}), \\ \text{and} \quad \max_{z\in\Omega}\max_{1\leq i\leq n} |\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_0^{\vartheta}| &\leq C(|\Omega|, \|\boldsymbol{A}^T\boldsymbol{A}\|, 0)(n^{\alpha(0)-1/2}). \end{aligned} \tag{52}$$

where C is some function depending on $|\Omega|$ and $\|\boldsymbol{A}^T\boldsymbol{A}\|$. From (**??**), the mini-batch gradient method update gives, for $t \geq 1$,

$$\begin{aligned} \boldsymbol{w}_t^{\vartheta} &= \boldsymbol{w}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{P}_{k+1}\boldsymbol{w}_k^{\vartheta} + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{x}_k^{\vartheta} \\ \text{and} \quad \boldsymbol{A}\boldsymbol{x}_t^{\vartheta} &= \boldsymbol{A}\boldsymbol{x}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{P}_{k+1}\boldsymbol{w}_k^{\vartheta} + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}, \end{aligned}$$

so by multiplying $\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)$ on both sides, we get

$$\begin{aligned} \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^{\vartheta} &= \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{P}_{k+1}\boldsymbol{w}_k^{\vartheta} \\ &\quad + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}, \\ \text{and} \quad \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^{\vartheta} &= \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_0^{\vartheta} + \sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{P}_{k+1}\boldsymbol{w}_k^{\vartheta} \\ &\quad + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}. \end{aligned} \tag{53}$$

Now assume the induction hypothesis, that is, for each $0 \leq t-1 < \vartheta$

$$\begin{aligned} &\max_{z\in\Omega}\max_{1\leq i\leq n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_{t-1}^{\vartheta}| \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t-1)\cdot n^{\alpha(t-1)-1/2},\\ \text{and}\quad &\max_{z\in\Omega}\max_{1\leq i\leq n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_{t-1}^{\vartheta}| \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t-1)\cdot n^{\alpha(t-1)-1/2}, \end{aligned} \tag{54}$$

with overwhelming probability. From (53), let $Y_{i,k} \overset{\text{def}}{=} \boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{P}_{k+1}(\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b})$. We will use Bernstein-type inequality. Note that $Y_{i,k}=\sum_{\ell\in B_k}X_{\ell,i,k}$ where $X_{\ell,i,k}\overset{\text{def}}{=}(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell(\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b})_\ell$. Observe,

$$\mu_k \overset{\text{def}}{=} \frac{1}{n}\sum_{\ell=1}^{n}X_{\ell,i,k} = \frac{1}{n}\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T\boldsymbol{w}_k^{\vartheta},$$

and

$$\sigma_k^2 \overset{\text{def}}{=} \frac{1}{n}\sum_{\ell=1}^{n}(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell^2(\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b})_\ell^2-\mu^2.$$

As for bounding μ, observe

$$\begin{aligned} R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T &= R(z;\boldsymbol{A}\boldsymbol{A}^T)(\boldsymbol{A}\boldsymbol{A}^T-z\mathbf{I}+z\mathbf{I})\\ &= \mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T). \end{aligned} \tag{55}$$

By left multiplying $\boldsymbol{e}_i^T$ and right multiplying $\boldsymbol{w}_k^{\vartheta}$, we get that

$$\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_k^{\vartheta} = \boldsymbol{e}_i^T\boldsymbol{w}_k^{\vartheta}+z\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_k^{\vartheta}$$

so that by using the induction hypothesis, with $z\in\Omega$, we have

$$\mu_k \leq \frac{1}{n}|\Omega|C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2} \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)n^{\alpha(k)-3/2},\quad \text{w.o.p.}$$

For the constant, we abuse notation so that $C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)=|\Omega|\max_{0\leq k\leq t-1}C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)$. We will use this sense for all constants below. As for σ^2, we have

$$\begin{aligned} \sigma_k^2 &\leq \frac{1}{n}\max_{\ell\neq i}\{(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell^2\}\|\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b}\|^2+\frac{1}{n}(\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_i^2|(\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b})_i|^2.\\ &= \frac{1}{n}\max_{\ell\neq i}\{(z\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_\ell)^2\}\|\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b}\|^2+\frac{1}{n}(1+z\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i)^2|(\boldsymbol{A}\boldsymbol{x}_k^{\vartheta}-\boldsymbol{b})_i|^2. \end{aligned} \tag{56}$$

The last equality follows from (55). By applying Assumption 3,

$$\max_{z\in\Omega}\max_{i\neq j}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_j\right| < n^{\alpha_0-1/2}\quad\text{and}\quad \max_{z\in\Omega}\max_{i}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i-\frac{1}{n}\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^T)\right| < n^{\alpha_0-1/2}, \tag{57}$$

together with the induction hypothesis and the definition of the stopping time ϑ, we have

$$\sigma_k^2 \leq C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t)\left[\frac{1}{n}(n^{(2\alpha_0-1)+2\theta})+\frac{1}{n}(n^{2\alpha(k)-1})\right] = C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t)n^{2\alpha(k)-2},\quad\text{w.o.p.}$$

Here we used that $\alpha_0+\theta<\alpha(k)$. Let $b_k \stackrel{\text{def}}{=} \max_{1\le\ell\le n} X_{l,i,k} = \max_{1\le\ell\le n}\{(e_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{A}^T)_\ell(\boldsymbol{A}\boldsymbol{x}_k^\vartheta - \boldsymbol{b})_\ell\}$. Again using (55), one equates b_k, as follows,

$$b_k = \max_{1\le\ell\le n}\{\big(\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_\ell\big)\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta\}. \tag{58}$$

We will bound $\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta$ using the induction hypothesis (54). For the other term, we look at two cases $\ell=i$ and $\ell\neq i$ and use Assumption 3:

$$\begin{aligned}
\ell\neq i \qquad & |\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_\ell| \le |\Omega|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_\ell \le |\Omega| n^{\alpha_0-1/2}\\
\ell=i \qquad & |\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_i| \le 1+|\Omega||\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i|\\
& \qquad\qquad 1+|\Omega|\big(|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{e}_i - \tfrac{1}{n}\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^T)| + \tfrac{1}{n}\operatorname{tr}(\boldsymbol{A}\boldsymbol{A}^T)\big)\\
& \qquad\qquad \le 1+|\Omega|(n^{\alpha_0-1/2}+\|\boldsymbol{A}^T\boldsymbol{A}\|).
\end{aligned}$$

Using this with the induction hypothesis (54) on $\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta$, we get a bound on b_k:

$$\begin{aligned}
b_k &= \max_{1\le\ell\le n}\{\big(\boldsymbol{e}_i^T(\mathbf{I}+zR(z;\boldsymbol{A}\boldsymbol{A}^T))\boldsymbol{e}_\ell\big)\boldsymbol{e}_\ell^T\boldsymbol{w}_k^\vartheta\}\\
&\le \big[|\Omega|n^{\alpha_0-1/2}+1+|\Omega|(n^{\alpha_0-1/2}+\|\boldsymbol{A}^T\boldsymbol{A}\|)\big]C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2}\\
&\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,\gamma,t)n^{\alpha(k)-1/2}, \quad \text{w.o.p.}
\end{aligned} \tag{59}$$

Note in the last inequality we used that $\alpha_0<1/2$ to conclude that $\alpha_0+\alpha(k)-1<\alpha(k)-1/2$.

Recall Proposition 2.5, that is,

$$\mathbb{P}\left(\frac{1}{\beta}\sum_{\ell\in B_k} X_{\ell,i,k}-\mu_k\ge\epsilon\right)\le\exp\left(-\frac{\beta\epsilon^2}{2\sigma_k^2+(2/3)(b_k-a)\epsilon}\right). \tag{60}$$

Let $\epsilon=n^{-3/2+\alpha(t)}$ in the above. Therefore after noting that β/n is the constant ζ, we deduce that the expression on the right-hand side inside the exponential is

$$-\frac{\beta\epsilon^2}{2\sigma^2+(2/3)(b-a)\epsilon}\lesssim\frac{-n\cdot n^{2(-3/2+\alpha(t))}}{n^{2\alpha(k)-2}+n^{\alpha(k)+\alpha(t)-2}}\lesssim -n^{\alpha(t)-\alpha(k)}. \tag{61}$$

Here we use $a\lesssim b$ to mean that there is a constant $C>0$ depending on $t,|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|$, but independent of n, such that $a\le Cb$. Since $\alpha(t)>\alpha(k)$ for all $k<t$, the probability in (60) goes down faster than any polynomial in n. Therefore, from (53) and the probability (60) with (61), we

have that

$$\begin{aligned}
&\max_{z\in\Omega}\max_{1\le i\le n}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_t^\vartheta\right| \\
&\qquad\le \max_{z\in\Omega}\max_{1\le i\le n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{w}_0^\vartheta| + \beta\sum_{k=0}^{t-1} c_{(t-1),k}[\mu_k+\epsilon] + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{x}_k^\vartheta \\
&\qquad\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,0)n^{\alpha(0)-1/2} + C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)\sum_{k=0}^{t-1} c_{(t-1),k}(n^{\alpha(k)-1/2}+n^{\alpha(t)-1/2}) \\
&\qquad\quad + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2} \\
&\qquad\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)n^{\alpha(t)-1/2},
\end{aligned} \tag{62}$$

with overwhelming probability. Here we used that $\alpha(k)<\alpha(t)$ for all $k<t$. Similarly, from (53) and the probability (60) with (61), one deduces, with overwhelming probability,

$$\begin{aligned}
&\max_{z\in\Omega}\max_{1\le i\le n}\left|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_t^\vartheta\right| \\
&\qquad\le \max_{z\in\Omega}\max_{1\le i\le n}|\boldsymbol{e}_i^T R(z;\boldsymbol{A}\boldsymbol{A}^T)\boldsymbol{A}\boldsymbol{x}_0^\vartheta| + \beta\sum_{k=0}^{t-1} c_{(t-1),k}[\mu_k+\epsilon] + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}\boldsymbol{A}\boldsymbol{x}_k^\vartheta \\
&\qquad\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,0)n^{\alpha(0)-1/2} + C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)\sum_{k=0}^{t-1} c_{(t-1),k}(n^{\alpha(k)-1/2}+n^{\alpha(t)-1/2}) \\
&\qquad\quad + \delta\zeta\sum_{k=0}^{t-1} c_{(t-1),k}C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,k)n^{\alpha(k)-1/2} \\
&\qquad\le C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)n^{\alpha(t)-1/2}.
\end{aligned} \tag{63}$$

The induction step is now shown and the result (51) follows for $0\le t\le\vartheta$ with $t\le T$.

For $T\ge t>\vartheta$, we have $\boldsymbol{w}_t^\vartheta=\boldsymbol{w}_{t+1}^\vartheta=\boldsymbol{w}_\vartheta^\vartheta$. We already showed that the result (51) holds for $(\boldsymbol{w}_0^\vartheta,\boldsymbol{A}\boldsymbol{x}_0^\vartheta)$ and $(\boldsymbol{w}_t^\vartheta,\boldsymbol{A}\boldsymbol{x}_t^\vartheta)$ where $t\le\vartheta$ and in particular when $t=\vartheta$. Thus, we immediately get that the result (51) holds for $T\ge t>\vartheta$. From (51), the desired proposition follows after noting that $\alpha(0)<\alpha(1)<\ldots<\alpha(T)\le\alpha$ and defining $C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,T)=\max_{0\le t\le T}C(|\Omega|,\|\boldsymbol{A}^T\boldsymbol{A}\|,t)$. ■

2.4.3. PROOF OF PROPOSITION 2.1

In this section, we will show that $\mathcal{E}_t^\nabla$ is small as $n\to\infty$.

Proof [Proof of Proposition 2.1] Recall for $t>0$, we have

$$\mathcal{E}_{t\wedge\vartheta}^\nabla=\sum_{k=1}^{t\wedge\vartheta}(\nabla\mathcal{R}(\boldsymbol{y}_t^\vartheta)^T\mathbf{N}_{(t\wedge\vartheta,k)}\left[\zeta\boldsymbol{A}^T\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^\vartheta-b\right)-\boldsymbol{A}^T\boldsymbol{P}_k\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^\vartheta-\boldsymbol{b}\right)\right] \tag{64}$$

We make use of Proposition 2.5, (proof is similar to Proposition 10 in [21]) by defining the following quantities:

$$X_I^{(t,k)}\overset{\text{def}}{=}-\nabla\mathcal{R}(\boldsymbol{y}_t^\vartheta)^T\mathbf{N}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{e}_I\boldsymbol{e}_I^T\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^\vartheta-\boldsymbol{b}\right), \tag{65}$$

where $\boldsymbol{e}_I$ denotes the I-th elementary basis that is included in the random batch B at the k-th iteration. Moreover, we define

$$\mu_{(t,k)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i\in[n]} \boldsymbol{X}_i^{(t,k)} = -\frac{1}{n}\sum_{i=1}^{n} \nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \mathbf{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_i \boldsymbol{e}_i^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right) \tag{66}$$

and hence

$$\mathcal{E}_{t\wedge\vartheta}^{\nabla} = \sum_{k=1}^{t\wedge\vartheta} \left(\sum_{i\in B_k} X_i^{(t,k)} - \beta\mu_{(t,k)} \right) \tag{67}$$

Since $k \le t \le T$, we apply Lemma 9 with an α chosen so that $\alpha' > \alpha > \alpha_0 + \theta$ to conclude that

$$\max_{1\le j\le n} \left|(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b})_j\right|^2 \le Cn^{2\alpha-1}, \quad \text{w.o.p.} \tag{68}$$

where the constant C depends on $\|\boldsymbol{A}^T\boldsymbol{A}\|, |\Omega|, T$, but it is independent of n and d. For each $k = 0, 1, 2 \ldots, t$, we can impose an upper bound on the empirical variance as follows:

$$\begin{aligned}
\sigma_{(t,k)}^2 &\stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{X}_i^{(t,k)}\right)^2 - \left(\mu_{(t,k)}\right)^2 \\
&\le \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{X}_i^{(t,k)}\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} \left(\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \mathbf{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_i\right)^2 \left(\boldsymbol{e}_i^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)\right)^2 \\
&\le \frac{1}{n} \max_{1\le j\le n} \left|\left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)_j\right|^2 \cdot ||\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})||^2 \cdot ||\mathbf{N}_{(t\wedge\vartheta,k)}||^2 \cdot ||\boldsymbol{A}||^2 \\
&\le C(t)n^{2\alpha-2} \quad \text{w.o.p.},
\end{aligned}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$, and time t (here take $\max$ of the constants over $1 \le k \le t$), but it is independent of n or d. All constants going forward will also have this property and we note that the constant will change from line to line. Similarly, using the same α as in (68), we can bound the following quantity:

$$\begin{aligned}
b_{(t,k)} = \max_{1\le i\le n} \boldsymbol{X}_i^{(t,k)} &\le \max_{1\le i\le n} \left|\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})^T \mathbf{N}_{(t\wedge\vartheta,k)} \boldsymbol{A}^T \boldsymbol{e}_i \boldsymbol{e}_i^T \left(\boldsymbol{A}\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)\right| \\
&\le \max_{1\le i\le n} |\left(A\boldsymbol{x}_{k-1}^{\vartheta} - \boldsymbol{b}\right)_i| \cdot ||\nabla\mathcal{R}(\boldsymbol{y}_t^{\vartheta})||_2 \cdot ||\mathbf{N}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T|| \\
&\le C(t)n^{\alpha-1/2} \quad \text{w.o.p.}
\end{aligned} \tag{69}$$

Applying Proposition 2.5 gives

$$\Pr\left(\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)} \ge \tilde{\varepsilon}\right) = \Pr\left(\frac{1}{\beta}\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \mu_{(t,k)} \ge \frac{\tilde{\varepsilon}}{\beta}\right) \le \exp\left(-\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma_{(t,k)}^2 + \frac{2}{3}(b_{(t,k)} - a)(\frac{\tilde{\varepsilon}}{\beta})}\right). \tag{70}$$

We first note that $\beta/n = \zeta$. Set $\tilde{\varepsilon} = \varepsilon/T$ with $\varepsilon = n^{\alpha'-1/2}$ and $a = 0$. Using the upper bounds on $b_{(t,k)}$ and $\sigma^2_{(t,k)}$,

$$\begin{aligned}\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma^2_{(t,k)} + \frac{2}{3}(b_{(t,k)} - a)(\frac{\tilde{\varepsilon}}{\beta})} &\geq C(t) \cdot \frac{T^2 n^{-1}\varepsilon^2}{n^{2\alpha-2} + Tn^{\alpha-1/2}n^{-1}\varepsilon} \geq C(t) \cdot \frac{T^2}{n^{2(\alpha-\alpha')} + Tn^{\alpha-\alpha'}} \\ &\geq C(T) \cdot \frac{T^2}{n^{2(\alpha-\alpha')} + Tn^{\alpha-\alpha'}} \underset{n\to\infty}{\to} \infty.\end{aligned} \tag{71}$$

Here again $C(T)$ is a positive constant independent of n and k. We note that we chose α so that $\alpha' > \alpha > \alpha_0 + \theta$. We can then lower bound $C(t)$ with $C(T)$ simply by letting $C(T) = \min_{1\leq t\leq T} C(t) > 0$. Hence

$$\Pr\left(\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)} \geq \tilde{\varepsilon}\right) \leq \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T \tag{72}$$

for some $c > 0$. Note that the constants c and $C(T)$ are independent of t and k, only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Similarly, by taking $-\boldsymbol{X}_i^{(t,k)}$ and using the same bounds on $b_{(t,k)}$ and $\sigma^2_{(t,k)}$, we get that

$$\Pr\left(-\big[\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)}\big] \geq \tilde{\varepsilon}\right) \leq \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T. \tag{73}$$

Therefore, it follows that

$$\Pr\left(\left|\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq \tilde{\varepsilon}\right) \leq 2\exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T, \tag{74}$$

for some $c > 0$ and $C(T) > 0$ where the constants do not depend on t, but do depend on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. We set $\varepsilon = n^{\alpha'-1/2}$. Applying the union bound twice and using (74), we get

$$\begin{aligned}\Pr\left(\max_{0\leq t\leq T} |\mathcal{E}^{\nabla}_{t\wedge\vartheta}| \geq \varepsilon\right) &\leq \sum_{t=0}^{T} \Pr\left(|\mathcal{E}^{\nabla}_{t\wedge\vartheta}| \geq \varepsilon\right) \\ &\leq \sum_{t=0}^{T} \Pr\left(\sum_{k=1}^{t\wedge\vartheta} \left|\left(\sum_{i\in B_k} X_i^{(t,k)} - \beta\mu_{(t,k)}\right)\right| \geq \varepsilon\right) \\ &\leq \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq \varepsilon/(t\wedge\vartheta)\right) \\ &\leq \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu_{(t,k)}\right| \geq \varepsilon/T\right) \\ &\leq T^2 \cdot \exp(-C(T)n^c)\end{aligned}$$

for some $c, C(T) > 0$. In the penultimate inequality, we used that $t\wedge\vartheta \leq T$. For the last inequality, we note that $\varepsilon/T = \tilde{\varepsilon}$ in (74) that the constants, $c, C(T) > 0$ in (74) hold for all $1 \leq t \leq T$. The result immediately follows.

■

2.4.4. PROOF OF PROPOSITION 2.2

Throughout this section, we define $\widetilde{\mathbf{N}}_{t,k} \overset{\text{def}}{=} \mathbf{N}_{t,k}(\boldsymbol{H})(\nabla^2\mathcal{R})\mathbf{N}_{t,k}(\boldsymbol{H})$ and we express the ij-th entry of $\widetilde{\mathbf{N}}_{t,k}$ as $\widetilde{N}_{ij}^{(t,k)}$. Moreover, we recall $\boldsymbol{w}_k \overset{\text{def}}{=} \boldsymbol{A}\boldsymbol{x}_k - \boldsymbol{b}$.

There will be require some substantial buildup of lemmas and propositions before we conclude with a proof of Proposition 2.2. The first of which is what we colloquially call the *key lemma*, based off the key lemma in [28, Lemma 14]. This version extends the result in [28], but maintains the essence of that lemma, in that, the on-diagonal entries of $\boldsymbol{A}p(\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{A}^T$ where p is a polynomial, self averages. The precise statement is below.

Lemma 10 (Key Lemma) *Fix $T > 0$ and suppose Assumption 5 holds. Let $p_k : \mathbb{C} \to \mathbb{C}$ be a k-degree polynomial with $k \leq T$ and coefficients which are independent of n and d. Then for some C depending on T and $|\Omega|$, the following holds with overwhelming probability*

$$\max_{0\leq k\leq T}\max_{1\leq \ell\leq n}\left|\boldsymbol{e}_\ell^T\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell - \tfrac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\right)\right| \leq n^{\alpha_0-1/2}C(T,|\Omega|)\|\mathcal{R}\|_{H^2}$$
$$\text{where} \quad \boldsymbol{W} \overset{\text{def}}{=} p_k(\boldsymbol{A}^T\boldsymbol{A})\cdot(\nabla^2\mathcal{R})\cdot p_k(\boldsymbol{A}^T\boldsymbol{A}).$$

Proof Define the matrix $\boldsymbol{K}(z,y) \overset{\text{def}}{=} \boldsymbol{A}R(z;\boldsymbol{A}^T\boldsymbol{A})(\nabla^2\mathcal{R})R(y;\boldsymbol{A}^T\boldsymbol{A})\boldsymbol{A}^T$. Let Ω by the contour and $\alpha_0 \in (0,\frac{1}{2})$ in Assumption 5. By Cauchy's integral formula, we can write

$$\begin{aligned}\boldsymbol{e}_\ell^T\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell &= \frac{-1}{(2\pi i)^2}\oint_\Omega \boldsymbol{e}_\ell^T p_k(z)p_k(y)\boldsymbol{K}(z,y)\boldsymbol{e}_\ell \,\mathrm{d}z\,\mathrm{d}y\\ \tfrac{1}{n}\operatorname{tr}(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T) &= \frac{-1}{(2\pi i)^2}\oint_\Omega p_k(z)p_k(y)\tfrac{1}{n}\operatorname{tr}\left(\boldsymbol{K}(z,y)\right)\,\mathrm{d}z\,\mathrm{d}y.\end{aligned} \tag{75}$$

Using Assumption 5, the following holds with overwhelming probability

$$\begin{aligned}|\boldsymbol{e}_\ell^T\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell - \tfrac{1}{n}\operatorname{tr}(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T)| &\leq \left|\frac{-1}{(2\pi i)^2}\oint_\Omega p_k(z)p_k(y)\left[\boldsymbol{e}_\ell^T\boldsymbol{K}(z,y)\boldsymbol{e}_\ell - \tfrac{1}{n}\operatorname{tr}(\boldsymbol{K}(z,y))\right]\,\mathrm{d}z\,\mathrm{d}y\right|\\ &\leq n^{\alpha_0-1/2}\frac{\|\mathcal{R}\|_{H^2}}{4\pi^2}\oint_\Omega |p_k(z)p_k(y)||\,\mathrm{d}z||\,\mathrm{d}y|\\ &\leq n^{\alpha_0-1/2}\frac{|\Omega|^2}{4\pi^2}\max_{z\in\Omega}|p_k(z)|^2\|\mathcal{R}\|_{H^2},\end{aligned} \tag{76}$$

Because the contour Ω is bounded, $\max_{0\leq k\leq T}\max_{z\in\Omega}|p_k(z)|$ is bounded independent of n and d, but it is dependent on T. Thus we get

$$\begin{aligned}\max_{0\leq k\leq T}\max_{1\leq\ell\leq n}|\boldsymbol{e}_\ell^T\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T\boldsymbol{e}_\ell - \tfrac{1}{n}\operatorname{tr}(\boldsymbol{A}\boldsymbol{W}\boldsymbol{A}^T)| &\leq \max_{0\leq k\leq T} n^{\alpha_0-1/2}\frac{|\Omega|^2}{4\pi^2}\max_{z\in\Omega}|p_k(z)|^2\|\mathcal{R}\|_{H^2}\\ &\leq n^{\alpha_0-1/2}C(T,|\Omega|)\|\mathcal{R}\|_{H^2},\end{aligned} \tag{77}$$

where the constant C depends on T and $|\Omega|$. The result immediately follows. ■

Next, we show that the error term $\mathcal{E}_t^{\nabla^2-\text{Diag}}$ is composed of multiple terms corresponding to four different types of errors: (1) the randomness in a quadratic form (see below $\mathcal{E}_t^{\nabla^2:\text{HW}}$), (2) the

randomness is solely due to the mini-batching ζ parameter (see below $\mathcal{E}_t^{(\nabla^2:\text{Z.1})}$ and $\mathcal{E}_t^{(\nabla^2:\text{Z.2})}$), (3). the randomness is linear (see below $\mathcal{E}_t^{(\nabla^2:\text{B.1})}$ and $\mathcal{E}_t^{(\nabla^2:\text{B.2})}$), and (4). the key lemma (see $\mathcal{E}_t^{\nabla^2:\text{KL}}$). These each will be handled accordingly. For instance, the quadratic form we will use Hanson-Wright concentration for martingales (see [1, Theorem 2.5], Lemma 19). For the linear martingale terms, we use Bernstein concentration proposed by Bardenet [5, Propsition 1.4] and restated in Proposition 2.5.

Lemma 11 *We can decompose the on-diagonal error term $\mathcal{E}_t^{\nabla^2\text{-Diag}}$ as follows*

$$\begin{aligned}\mathcal{E}_t^{\nabla^2\text{-Diag}} &\overset{def}{=} \frac{1}{2}\sum_{k=1}^{t} \mathring{\boldsymbol{M}}_k^T \widetilde{\mathbf{N}}_{t,k} \mathring{\boldsymbol{M}}_k - \frac{\gamma^2}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n} \frac{(\zeta-\zeta^2)}{n} \operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{t,k}\boldsymbol{A}^T\right) \boldsymbol{w}_{k-1,\ell}^2 \\ &= \mathcal{E}_t^{(\nabla^2:\text{KL})} + \mathcal{E}_t^{(\nabla^2:\text{Z.1})} + \mathcal{E}_t^{(\nabla^2:\text{Z.2})} + \mathcal{E}_t^{(\nabla^2:\text{B.1})} + \mathcal{E}_t^{(\nabla^2:\text{B.2})} + \mathcal{E}_t^{(\nabla^2:\text{HW})}\end{aligned} \tag{78}$$

where the six error terms are

$$\begin{aligned}
\mathcal{E}_t^{(\nabla^2:\text{KL})} &= \frac{1}{2}(\zeta-\zeta^2)\sum_{k=1}^{t}\sum_{ij}\sum_{\ell=1}^{n}\left(A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\widetilde{N}_{ij}^{(t,k)}\right) - \frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2 \\
\mathcal{E}_t^{(\nabla^2:\text{Z.1})} &= \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} \\
\mathcal{E}_t^{(\nabla^2:\text{Z.2})} &= \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\sum_{\ell=1}^{n}\left(\zeta^2-\frac{\beta(\beta-1)}{n(n-1)}\right)A_{\ell j}A_{\ell i}\widetilde{N}_{ij}^{(t,k)}w_{k-1,\ell}^2 \\
\mathcal{E}_t^{(\nabla^2:\text{B.1})} &= \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left[\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} - \zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right] \\
\mathcal{E}_t^{(\nabla^2:\text{B.2})} &= \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left[\zeta^2\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} - \zeta\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right] \\
\mathcal{E}_t^{(\nabla^2:\text{HW})} &= \frac{1}{2}\sum_{k=1}^{t}\sum_{ij}\widetilde{N}_{ij}^{(t,k)}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1} - \mathbb{E}[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_t\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}]\right].
\end{aligned} \tag{79}$$

We indexed the error terms by the acronyms inspired by the results applied to bound them: Key Lemma (KL), Zeta (Z), Bardenet (B), and Hanson-Wright (HW).

First, we state a result first proven in in [28, Lemma B.1] that will be used in the proof of Lemma 11.

Lemma 12 (Lemma B.1 [28]) *Suppose that $\boldsymbol{u}$ and $\boldsymbol{v}$ are fixed vectors in $\mathbb{R}^n$. Then*

$$\mathbb{E}\left[\left(\sum_{i\in B}u_iv_i\right)^2\right] = \frac{\beta}{n}\frac{\beta-1}{n-1}(\boldsymbol{u}^T\boldsymbol{v})^2 + \left(\frac{\beta}{n}-\frac{\beta}{n}\frac{\beta-1}{n-1}\right)\sum_{i=1}^{n}(u_iv_i)^2.$$

We are now ready to prove Lemma 11.

Proof [Proof of Lemma 11] Observe that

$$\frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\widetilde{\mathbf{N}}_{t,k}\mathring{\boldsymbol{M}}_k=\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\mathring{\boldsymbol{M}}_{k,i}\widetilde{N}_{ij}^{(t,k)}\mathring{\boldsymbol{M}}_{k,j} \tag{80}$$

and that the product of the entries of the Martingale increments defined in (20) gives us

$$\begin{aligned}
\mathring{\boldsymbol{M}}_{k,i}\mathring{\boldsymbol{M}}_{k,j}&=\left[\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\right]\left[\zeta\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\right]\\
&=\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\\
&\quad-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}+\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\\
&=\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-2\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\\
&\quad+\left(\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)\\
&\quad+\left(\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\right)\\
&\quad+\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\\
&=-\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}+\mathcal{E}_k^{(\nabla^2,1)}(i,j)+\mathcal{E}_k^{(\nabla^2,2)}(i,j)+\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}
\end{aligned} \tag{81}$$

where we introduce two error terms

$$\begin{aligned}
\mathcal{E}_k^{(\nabla^2,1)}(i,j)&\stackrel{\text{def}}{=}\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\\
\mathcal{E}_k^{(\nabla^2,2)}(i,j)&\stackrel{\text{def}}{=}\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}.
\end{aligned} \tag{82}$$

We note that the conditional expectation of $\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}$ is precisely $\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}$ (same for the other term). We will show, in fact, that $\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}$ will concentrate around its mean.

We now consider the term $\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}$ in (81) and we perform a Doob's decomposition on this term, that is,

$$\begin{aligned}
\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}&=\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}-\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]\\
&\quad+\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right].
\end{aligned} \tag{83}$$

The first term in the inequality will be small as $\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}$ will concentrate around its mean. It remains to simplify the conditional expectation term, $\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]$, which we do so below. Noting that $\boldsymbol{P}_k=\sum_{m\in B_{k-1}}\boldsymbol{e}_m\boldsymbol{e}_m^T$,

$$\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]=\mathbb{E}\left[\left(\sum_{m\in B_k}A_{mi}w_{k-1,m}\right)\left(\sum_{\ell\in B_k}A_{\ell j}w_{k-1,\ell}\right)|\mathcal{F}_{k-1}\right].$$

We want to apply Lemma 12 to the above, but it is not in the form that Lemma 12. To get it into the correct form, we use polarization:

$$\mathbb{E}\left[\Big(\sum_{m\in B_{k-1}} A_{mi}w_{k-1,m}\Big)\Big(\sum_{\ell\in B_{k-1}} A_{\ell j}w_{k-1,\ell}\Big)|\mathcal{F}_{k-1}\right] = \frac{1}{4}\,\mathbb{E}\left[\Big(\sum_{m\in B_{k-1}} A_{mj}w_{k-1,m} + A_{mi}w_{k-1,m}\Big)^2|\mathcal{F}_{k-1}\right] - \frac{1}{4}\,\mathbb{E}\left[\Big(\sum_{m\in B_{k-1}} A_{mj}w_{k-1,m} - A_{mi}w_{k-1,m}\Big)^2|\mathcal{F}_{k-1}\right]. \tag{84}$$

Now we apply Lemma 12 to each term in (84) and thus, after simplifying,

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right] &= \frac{1}{4}\,\mathbb{E}\left[\Big(\sum_{m\in B_{k-1}} A_{mj}w_{k-1,m} + A_{mi}w_{k-1,m}\Big)^2|\mathcal{F}_{k-1}\right] \\
&\quad - \frac{1}{4}\,\mathbb{E}\left[\Big(\sum_{m\in B_{k-1}} A_{mj}w_{k-1,m} - A_{mi}w_{k-1,m}\Big)^2|\mathcal{F}_{k-1}\right] \\
&= \frac{1}{4}\left[\frac{\beta(\beta-1)}{n(n-1)}\left[(\boldsymbol{A}\boldsymbol{e}_j+\boldsymbol{A}\boldsymbol{e}_i)^T\boldsymbol{w}_{k-1}\right]^2 + \left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{\ell=1}^n (A_{\ell j}+A_{\ell i})^2 w_{k-1,\ell}^2\right] \\
&\quad - \frac{1}{4}\left[\frac{\beta(\beta-1)}{n(n-1)}\left[(\boldsymbol{A}\boldsymbol{e}_j-\boldsymbol{A}\boldsymbol{e}_i)^T\boldsymbol{w}_{k-1}\right]^2 + \left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{\ell=1}^n (A_{\ell j}-A_{\ell i})^2 w_{k-1,\ell}^2\right] \\
&= \left(\frac{\beta(\beta-1)}{n(n-1)}\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} + \left(\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2 \\
&= \zeta^2\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} + (\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2 \\
&\quad + \left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} + \left(\left[\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right]-(\zeta-\zeta^2)\right)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2.
\end{aligned} \tag{85}$$

In the last equality, we added and subtracted terms corresponding to $\zeta \approx \frac{\beta-1}{n-1}$ when n is large. Plugging in (85) into (81) we obtain

$$\begin{aligned}
\mathring{\boldsymbol{M}}_{k,i}\mathring{\boldsymbol{M}}_{k,j} &= (\zeta-\zeta^2)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2 + \left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} \\
&\quad + \left(\left[\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right]-(\zeta-\zeta^2)\right)\sum_{\ell=1}^n A_{\ell j}A_{\ell i}w_{k-1,\ell}^2 + \mathcal{E}_k^{(\nabla^2,1)}(i,j) + \mathcal{E}_k^{(\nabla^2,2)}(i,j) \\
&\quad + \left(\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1} - \mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]\right),
\end{aligned} \tag{86}$$

where $\mathcal{E}_k^{(\nabla^2,1)}(i,j)$ and $\mathcal{E}_k^{(\nabla^2,2)}(i,j)$ are defined in (82). Returning to (80) using the martingale increment computation (86),

$$
\begin{aligned}
&\frac{1}{2}\sum_{k=1}^{t}\mathring{\boldsymbol{M}}_k^T\widetilde{\mathbf{N}}_{t,k}\mathring{\boldsymbol{M}}_k-\frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\\
&=\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\mathring{\boldsymbol{M}}_{k,i}\widetilde{N}_{ij}^{(t,k)}\mathring{\boldsymbol{M}}_{k,j}-\frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\\
&=\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}(\zeta-\zeta^2)\sum_{\ell=1}^{n}A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\widetilde{N}_{ij}^{(t,k)}-\frac{1}{2}\sum_{k=1}^{t}\sum_{\ell=1}^{n}\frac{(\zeta-\zeta^2)}{n}\operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{t,k}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left(\zeta^2-\frac{\beta(\beta-1)}{n(n-1)}\right)\widetilde{N}_{ij}^{(t,k)}\sum_{\ell=1}^{n}A_{\ell j}A_{\ell i}w_{k-1,\ell}^2\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\left[\mathcal{E}_k^{(\nabla^2,1)}(i,j)+\mathcal{E}_k^{(\nabla^2,2)}(i,j)\right]\\
&+\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}-\mathbb{E}\left[\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}|\mathcal{F}_{k-1}\right]\right].
\end{aligned}
\tag{87}
$$

Using that $\zeta=\frac{\beta}{n}$, the result follows by matching up terms in (87) with (79) and using the definitions

$$
\mathcal{E}_t^{(\nabla^2:\mathrm{B}.1)}\overset{\text{def}}{=}\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\mathcal{E}_k^{(\nabla^2,1)}(i,j)\quad\text{and}\quad\mathcal{E}_t^{(\nabla^2:\mathrm{B}.2)}\overset{\text{def}}{=}\frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\widetilde{N}_{ij}^{(t,k)}\mathcal{E}_k^{(\nabla^2,2)}(i,j).
$$

■

We now show that each of the terms in (87) are small with overwhelming probability.

2.4.5. CONTROLLING $\mathcal{E}_t^{(\nabla^2:\mathrm{KL})}$

The following lemma controls the error term using the key lemma, Lemma 10.

Lemma 13 *Fix $T>0$. For any $\alpha>\alpha_0+\theta$,*

$$
\max_{0\le t\le T}\left|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:KL)}\right|\le C(T)n^{\alpha_0-1/2+2\theta}
\tag{88}
$$

holds with overwhelming probability.

Proof By applying Lemma 10 (Key Lemma) and Lemma 9 we get

$$\begin{aligned}\max_{0\le t\le T}\left|\mathcal{E}_t^{(\nabla^2:\mathrm{KL})}\right| &= \frac{1}{2}(\zeta-\zeta^2)\max_{0\le t\le T}\left|\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\left(\boldsymbol{e}_\ell^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{e}_\ell w_{k-1,\ell}^2\right)-\sum_{k=1}^{t\wedge\vartheta}\sum_{\ell=1}^{n}\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\right)w_{k-1,\ell}^2\right| \\ &\le \frac{1}{2}(\zeta-\zeta^2)\cdot T\cdot\max_{0\le t\le T}\max_{1\le\ell\le n}\left|\boldsymbol{e}_\ell^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{e}_\ell-\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\right)\right|\cdot\|\boldsymbol{w}_{k-1}^{\vartheta}\|^2 \\ &\le C(T)n^{\alpha_0-1/2}\cdot n^{2\theta}\end{aligned} \tag{89}$$

where $C(T)$ depends on T, $|\Omega|$ and γ^2 but independent of n and d. ■

2.4.6. CONTROLLING $\mathcal{E}_t^{\nabla^2:\mathrm{B.1}}$, $\mathcal{E}_t^{\nabla^2:\mathrm{B.2}}$

For the terms $\mathcal{E}_t^{\nabla^2:\mathrm{B.1}}$ and $\mathcal{E}_t^{\nabla^2:\mathrm{B.2}}$, we use Proposition 2.5 as the randomness due to mini-batching, $\boldsymbol{P}_k$, is linear in these terms.

Lemma 14 *Fix $T>0$. Let $\alpha'>\alpha_0+\theta$ and $\alpha>0$ such that $\alpha'>\alpha>\alpha_0+\theta$. The following holds with overwhelming probability*

$$\max_{0\le t\le T}|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{B.1})}|\le n^{\alpha'-1/2}, \tag{90}$$

provided $\theta<\alpha'-\alpha$.

Proof Recall for $t>0$, we have

$$\begin{aligned}\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{B.1})} &= \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{i,j}\left[\zeta^2\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t\wedge\vartheta,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{t-1}\widetilde{N}_{ij}^{(t\wedge\vartheta,k)}\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right] \\ &= \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left[\zeta^2\boldsymbol{w}_{k-1}^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}-\zeta\boldsymbol{w}_{k-1}^T\boldsymbol{P}_k\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right].\end{aligned}$$

Proceeding similarly to Proposition 2.1 for $\mathcal{E}_t^{\nabla}$, we define

$$X_I^{(t,k)}\overset{\text{def}}{=}-\frac{\zeta}{2}(\boldsymbol{w}_{k-1}^{\vartheta})^T\boldsymbol{e}_I\boldsymbol{e}_I^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\quad\text{and}\quad\mu^{(t,k)}\overset{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}X_i^{(t,k)}, \tag{91}$$

and we observe that

$$\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{B.1})}=\sum_{k=1}^{t\wedge\vartheta}\left(\sum_{i\in B_{k-1}}X_i^{(t,k)}-\beta\mu^{(t,k)}\right).$$

Before proceeding, we note that $\|\widetilde{\mathbf{N}}_{t,k}\|$ for $t\ge k$ is bounded by a constant depending on $\|\boldsymbol{A}^T\boldsymbol{A}\|$ and t, but independent of n and d. We apply Lemma 9 with α satisfying $\alpha'>\alpha>\alpha_0+\theta$ and get

that $\max_{1\le j\le n}|w_{k-1,j}| \le n^{\alpha-1/2}$. Using the definition of ϑ, we have the bounds

$$\begin{aligned}
\sigma^2_{(t,k)} &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i^{(t,k)}\right)^2 - \left(\mu^{(t,k)}\right)^2 \\
&\le \frac{\zeta^2}{4n}\sum_{i=1}^{n}\left(\boldsymbol{e}_i^T\boldsymbol{w}_{k-1}^{\vartheta}\right)^2\left(\boldsymbol{e}_i^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\right)^2 \\
&\le \frac{\zeta^2}{4n}\max_{1\le j\le n}|w_{k-1,j}^{\vartheta}|^2\cdot||\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}||^2 \\
&\le \frac{\zeta^2}{4n}\max_{1\le j\le n}|w_{k-1,j}^{\vartheta}|^2||\boldsymbol{w}_{k-1}^{\vartheta}||^2||\boldsymbol{A}||^4\cdot||\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}||^2 \\
&\le C(t)n^{2(\alpha+\theta-1)} \qquad \text{w.o.p,}
\end{aligned}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$, and time t (here take $\max$ of the constants over $1 \le k \le t$), but it is independent of n or d. All constants going forward will also have this property and we note that the constant will change from line to line.

Similarly, applying Lemma 9 with α, we can bound the following quantity:

$$\begin{aligned}
b &= \max_{1\le i\le n} X_i^{(t,k)} \\
&\le \max_{1\le i\le n}|\frac{\zeta}{2}(\boldsymbol{w}_{k-1}^{\vartheta})^T\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}| \\
&\le \frac{\zeta}{2}\max_{1\le i\le n}|w_{k-1,i}^{\vartheta}|\cdot||\boldsymbol{e}_i^T\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}||\cdot||\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}|| \\
&\le \frac{\zeta}{2}\left(\max_{1\le i\le n}|w_{k-1,i}|\right)\cdot||\boldsymbol{w}_{k-1}^{\vartheta}||\cdot\|\boldsymbol{A}\|^2\cdot\|\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\| \\
&\le C(t)n^{\alpha-\frac{1}{2}+\theta} \qquad \text{w.o.p.}
\end{aligned} \tag{92}$$

Applying Proposition 2.5 gives

$$\begin{aligned}
\Pr\left(\sum_{i=1}^{\beta}X_i^{(t,k)} - \beta\mu^{(t,k)} \ge \tilde{\varepsilon}\right) &= \Pr\left(\frac{1}{\beta}\sum_{i=1}^{\beta}X_i^{(t,k)} - \mu^{(t,k)} \ge \frac{\tilde{\varepsilon}}{\beta}\right) \\
&\le \exp\left(-\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma^2_{(t,k)} + \frac{2}{3}(b-a)(\frac{\tilde{\varepsilon}}{\beta})}\right).
\end{aligned} \tag{93}$$

We note that $\beta/n = \zeta$. Set $\tilde{\varepsilon} = \varepsilon/T$ with $\varepsilon = n^{\alpha'-1/2}$ where $\alpha' > \alpha$ and a = 0. Using the upper bounds on b and $\sigma^2_{(t,k)}$,

$$\frac{\beta(\frac{\tilde{\varepsilon}}{\beta})^2}{2\sigma^2_{(t,k)} + \frac{2}{3}(b-a)(\frac{\tilde{\varepsilon}}{\beta})} \ge C(t)\cdot\frac{T^{-2}n^{-1}\varepsilon^2}{n^{2(\alpha+\theta-1)} + T^{-1}n^{\alpha-1/2+\theta}n^{-1}\varepsilon} \ge C(t)\cdot\frac{T^{-2}}{n^{2(\alpha-\alpha'+\theta)} + T^{-1}n^{\alpha-\alpha'+\theta}}.$$

Here again $C(t)$ is a positive constant independent of n and d. By the choice of $\alpha' > \alpha$ and $\alpha' - \alpha > \theta$, we have that the right-hand-side goes to infinity. We can then lower bound $C(t)$ with

$C(T)$ simply by letting $C(T) = \min_{1\le t\le T} C(t) > 0$. Hence

$$\Pr\left(\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu^{(t,k)} \ge \tilde{\varepsilon}\right) \le \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T \tag{94}$$

for some $c > 0$. Note that the constants c and $C(T)$ are independent of t and k, only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Similarly, by taking $-\boldsymbol{X}_i^{(t,k)}$ and using the same bounds on b and $\sigma^2_{(t,k)}$, we get that

$$\Pr\left(-\big[\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu^{(t,k)}\big] \ge \tilde{\varepsilon}\right) \le \exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T. \tag{95}$$

Therefore, it follows that

$$\Pr\left(\left|\sum_{i=1}^{\beta} \boldsymbol{X}_i^{(t,k)} - \beta\mu^{(t,k)}\right| \ge \tilde{\varepsilon}\right) \le 2\exp\left(-C(T)n^c\right) \quad \text{when } \tilde{\varepsilon} = n^{\alpha'-1/2}/T, \tag{96}$$

for some $c > 0$ and $C(T) > 0$ where the constants do not depend on t and n and d. The constants do depend on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. We set $\varepsilon = n^{\alpha'-1/2}$. Applying the union bound twice and using (96), we get

$$\begin{aligned}
\Pr\left(\max_{0\le t\le T} |\mathcal{E}^{(\nabla^2:\text{B.1})}_{t\wedge\vartheta}| \ge \varepsilon\right) &\le \sum_{t=0}^{T} \Pr\left(|\mathcal{E}^{(\nabla^2:\text{B.1})}_{t\wedge\vartheta}| \ge \varepsilon\right) \\
&\le \sum_{t=0}^{T} \Pr\left(\sum_{k=1}^{t\wedge\vartheta} \left|\left(\sum_{i\in B_k} X_i^{(t,k)} - \beta\mu^{(t,k)}\right)\right| \ge \varepsilon\right) \\
&\le \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu^{(t,k)}\right| \ge \varepsilon/(t\wedge\vartheta)\right) \\
&\le \sum_{t=0}^{T}\sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left|\sum_{i=1}^{\beta} X_i^{(t,k)} - \beta\mu^{(t,k)}\right| \ge \varepsilon/T\right) \\
&\le T^2 \cdot \exp(-C(T)n^c)
\end{aligned}$$

for some $c, C(T) > 0$. In the penultimate inequality, we used that $t\wedge\vartheta \le T$. For the last inequality, we note that $\varepsilon/T = \tilde{\varepsilon}$ in (96) and that the constants, $c, C(T) > 0$ in (96) hold for all $1 \le t \le T$. The result immediately follows. ■

An immediate corollary of the result is that $\mathcal{E}_t^{\nabla^2:\text{B.2}}$ is also small.

Corollary 15 *Fix $T > 0$. Let $\alpha' > \alpha_0 + \theta$ and $\alpha > 0$ such that $\alpha' > \alpha > \alpha_0 + \theta$. The following holds with overwhelming probability*

$$\max_{0\le t\le T\wedge\vartheta} |\mathcal{E}^{(\nabla^2:\text{B.2})}_{t\wedge\vartheta}| = C(T)n^{\alpha'-1/2}, \tag{97}$$

provided $\alpha' - \alpha > \theta$.

Proof We observe that for $t \geq 0$ we have

$$\begin{aligned}
\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{B.2})} &= \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\sum_{i,j}\left[\zeta^2 \boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1} - \zeta\boldsymbol{e}_j^T\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\widetilde{N}_{ij}^{(t,k)}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\sum_{i,j}\left[\zeta^2\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)_i\widetilde{N}_{ij}^{(t,k)}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)_j - \zeta\left(\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\right)_j\widetilde{N}_{ij}^{(t,k)}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)_i\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\left[\zeta^2\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)^T\widetilde{\mathbf{N}}_{t,k}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right) - \zeta\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)^T\widetilde{\mathbf{N}}_{t,k}\left(\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\right)\right] \\
&= \frac{1}{2}\sum_{k=1}^{t}\left[\zeta^2\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)^T\widetilde{\mathbf{N}}_{t,k}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right) - \zeta\left(\boldsymbol{A}^T\boldsymbol{P}_k\boldsymbol{w}_{k-1}\right)^T\widetilde{\mathbf{N}}_{t,k}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}\right)\right] \\
&= \mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{B.1})}
\end{aligned}$$

where in the penultimate step we used the fact that $\widetilde{\mathbf{N}}_{(t,k)}$ is a symmetric matrix for each $t \in [T\wedge\vartheta]$. By applying Proposition 14 we achieve our desired result. ■

2.4.7. CONTROLLING $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.1})}$, $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}$

The zeta errors, $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.1})}$, $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}$, arise purely from an approximation of ζ^2 with $\frac{\beta(\beta-1)}{n(n-1)}$. As such, it is simple to show these terms indeed vanish in n.

Lemma 16 *The following holds*

$$\max_{0\leq t\leq T}|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.1})}| \leq n^{2\theta-1} \quad \textit{and} \quad \max_{0\leq t\leq T}|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}| \leq C(T)n^{2\theta-1}. \tag{98}$$

Proof A simple computation shows that

$$\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 = \frac{n\beta(\beta-1)-\beta^2(n-1)}{n^2(n-1)} = \frac{\zeta(\zeta-1)}{n-1} \leq C(\zeta)n^{-1}. \tag{99}$$

First we show the result for $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.1})}$ (see (79)). By applying the definition of the stopping time ϑ, we deduce that

$$\begin{aligned}
|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.1})}| &\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}|\sum_{i,j}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\right)_i\widetilde{N}_{ij}^{(t,k)}\left(\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}\right)_j| \\
&= \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}|\left(\boldsymbol{A}^T\boldsymbol{w}_{t-1}^{\vartheta}\right)^T\widetilde{\mathbf{N}}_{t,k}\boldsymbol{A}^T\boldsymbol{w}_{k-1}^{\vartheta}| \\
&\leq \frac{\zeta(\zeta-1)}{n-1}\cdot\frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}||\boldsymbol{A}||^2\cdot||\widetilde{\mathbf{N}}_{t,k}||\cdot||\boldsymbol{w}_{k-1}^{\vartheta}||^2 \\
&\leq C(T)n^{2\theta-1},
\end{aligned}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T$, and $|\Omega|$, (here take $\max$ of the constants over $1 \leq k \leq T$), but it is independent of n or d. Now taking the maximum over $0 \leq t \leq T$ proves the result.

Next we show the result holds for $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}$ (see (79)). Applying Lemma (10) with $\boldsymbol{W} = \widetilde{\mathbf{N}}_{t,k}$, we deduce that

$$
\begin{aligned}
|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\text{Z.2})}| &\leq \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \left| \sum_{i,j} \sum_{\ell=1}^{n} \boldsymbol{A}_{\ell j} \boldsymbol{A}_{\ell i} \widetilde{N}_{ij}^{(k,t)} (\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2 \right| \\
&= \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \left| \sum_{\ell=1}^{n} (\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T)_{\ell\ell} (\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2 \right| \\
&= \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \left| \sum_{\ell=1}^{n} \left[(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T)_{\ell\ell} - \frac{1}{n} \operatorname{tr}\left(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T \right) + \frac{1}{n} \operatorname{tr}(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T) \right] (\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2 \right| \\
&\leq \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \left| \sum_{\ell=1}^{n} \left[(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T)_{\ell\ell} - \frac{1}{n} \operatorname{tr}\left(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T \right) \right] (\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2 \right| \\
&\quad + \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \left| \sum_{\ell=1}^{n} \frac{1}{n} \operatorname{tr}(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T) (\boldsymbol{w}_{k-1,\ell}^{\vartheta})^2 \right| \\
&\leq \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \max_{1\leq\ell\leq n} |(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T)_{\ell\ell} - \frac{1}{n} \operatorname{tr}\left(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T \right)| \, ||\boldsymbol{w}_{k-1}^{\vartheta}||_2^2 \\
&\quad + \frac{\zeta(\zeta-1)}{n-1} \cdot \frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \frac{1}{n} \operatorname{tr}\left(\boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T \right) ||\boldsymbol{w}_{k-1}^{\vartheta}||_2^2 \\
&\leq C'(t) \cdot n^{\alpha_0 - 1/2} \cdot n^{2\theta-1} + C(t) n^{2\theta-1},
\end{aligned}
$$

where the constants C and C' are dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, T, |\Omega|$, and time t (here take $\max$ of the constants over $1 \leq k \leq t$), but it is independent of n or d. Since $0 \leq t \leq T$ and t is finite, we can take the maximum over both sides. ■

2.4.8. CONTROLLING $\mathcal{E}_t^{(\nabla^2:\text{HW})}$: HANSON-WRIGHT

Lastly, we need to control the term,

$$
\begin{aligned}
\mathcal{E}_t^{(\nabla^2:\text{HW})} &= \frac{1}{2} \sum_{k=1}^{t} \sum_{ij} \widetilde{N}_{ij}^{(t,k)} \left[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} - \mathbb{E}[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \boldsymbol{e}_j^T \boldsymbol{A}^T \boldsymbol{P}_t \boldsymbol{w}_{k-1} | \mathcal{F}_{k-1}] \right] \\
&= \frac{1}{2} \sum_{k=1}^{t} \left[(\boldsymbol{P}_k \boldsymbol{w}_{k-1})^T \boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T (\boldsymbol{P}_k \boldsymbol{w}_{k-1}) - \mathbb{E}\left[(\boldsymbol{P}_k \boldsymbol{w}_{k-1})^T \boldsymbol{A} \widetilde{\mathbf{N}}_{t,k} \boldsymbol{A}^T (\boldsymbol{P}_k \boldsymbol{w}_{k-1}) | \mathcal{F}_{k-1} \right] \right].
\end{aligned} \tag{100}
$$

Unlike the previous terms, this term has the randomness induced from mini-batching, $\boldsymbol{P}_k$, sandwiching a matrix. As such, we can not apply Proposition 2.5 due to the quadratic form. Instead, we use Hanson-Wright concentration result for quadratic forms.

Before continuing, we introduce some definitions, lemmas, and remarks related to Hanson-Wright concentration of quadratic forms.

Definition 17 (Convex concentration property, [1]) *Let $\boldsymbol{X}$ be a random vector in $\mathbb{R}^n$. We will say that $\boldsymbol{X}$ has the convex concentration property with constant K if for every $1-$Lipschitz convex function $\varphi : \mathbb{R}^n \to \mathbb{R}$. we have $\mathbb{E}[\varphi(\boldsymbol{X})] < \infty$ and for every $t > 0$,*

$$\Pr(|\varphi(\boldsymbol{X}) - \mathbb{E}\varphi(\boldsymbol{X})| \geq t) \leq 2\exp(-t^2/K^2).$$

Remark 18 Let x_i be an entry of $\boldsymbol{X}$. By a simple scaling, the previous remark can extend to $x_1, \cdots, x_n \in [a, b]$, in which case K in the definition above will be replaced by $K(b-a)$.

What is interesting for us is that vectors obtained via sampling without replacement follow the convex concentration property ([1, Remark 2.3]). More precisely, if $x_1, \cdots, x_n \in [0, 1]$ and the random vector $\boldsymbol{X} = (X_1, \cdots, X_m)$ with $m \leq n$ is obtained by sampling without replacement m numbers from the set $\{x_1, \cdots, x_n\}$, then $\boldsymbol{X}$ satisfies the convex concentration property with an absolute constant K. In this sense, the following lemma ([1, Theorem 2.5]) will be useful to us.

Lemma 19 (Hanson-Wright concentration for sampling without replacement, Theorem 2.5 [1]) *Let $\boldsymbol{X}$ be a mean zero random vector in $\mathbb{R}^n$. If $\boldsymbol{X}$ has the convex concentration property with constant K, then for any $n \times n$ matrix $\boldsymbol{A}$ and every $t > 0$,*

$$\Pr(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}| \geq t) \leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{K^2\|\boldsymbol{A}\|}\right)\right),$$

for some universal constant C.

Remark 20 The assumption that $\boldsymbol{X}$ is centered is introduced just to simplify the statement of the theorem. Note that if $\boldsymbol{X}$ has the convex concentration property with constant K, then so does $\tilde{\boldsymbol{X}} = \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}]$. Moreover, observe,

$$\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} = (\tilde{\boldsymbol{X}} + \mathbb{E}[\boldsymbol{X}])^T\boldsymbol{A}(\tilde{\boldsymbol{X}} + \mathbb{E}[\boldsymbol{X}]) = \tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} + 2\tilde{\boldsymbol{X}}^T\boldsymbol{A}\,\mathbb{E}[\boldsymbol{X}] + \mathbb{E}[\boldsymbol{X}]^T\boldsymbol{A}\,\mathbb{E}[\boldsymbol{X}]$$

$$\text{and} \quad \mathbb{E}\left[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\right] = \mathbb{E}\left[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}}\right] + \mathbb{E}\left[\boldsymbol{X}\right]^T\boldsymbol{A}\mathbb{E}\left[\boldsymbol{X}\right],$$

as $\mathbb{E}\left[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\mathbb{E}\left[\boldsymbol{X}\right]\right] = 0$. This implies by Lemma 19 and convex concentration property for linear functions,

$$\begin{aligned}
&\Pr(|\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} - \mathbb{E}[\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}]| \geq t) \\
&\quad\leq \Pr(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}} - \mathbb{E}[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\tilde{\boldsymbol{X}}]| \geq t/3) + 2\Pr(|\tilde{\boldsymbol{X}}^T\boldsymbol{A}\,\mathbb{E}[\boldsymbol{X}] - \mathbb{E}\left[\tilde{\boldsymbol{X}}^T\boldsymbol{A}\,\mathbb{E}[\boldsymbol{X}]\right]| \geq t/3) \\
&\quad\leq 2\exp\left(-\frac{1}{C}\min\left(\frac{t^2}{2\cdot 9K^4\|\boldsymbol{A}\|_{HS}^2}, \frac{t}{3K^2\|\boldsymbol{A}\|}\right)\right) + 2\cdot 2\exp\left(-\frac{t^2}{9K^2\|\boldsymbol{A}\,\mathbb{E}[\boldsymbol{X}]\|_2^2}\right).
\end{aligned}$$

Using Lemma 19 and Remark 20 we can show $\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})}$ is small for large n and d.

Lemma 21 *For all α' and α such that $\alpha' > \alpha > \alpha_0 + \theta$ and θ, we have*

$$\max_{1\leq t\leq T}|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:HW)}| \leq C(T)n^{2\alpha'-1/2} \quad \textit{w.o.p.},$$

provided $0 < \theta < 2\alpha' - \alpha$.

Proof Let $\boldsymbol{X}_k \stackrel{\text{def}}{=} \boldsymbol{P}_k \boldsymbol{w}_{k-1}^{\vartheta}$ and $\boldsymbol{D}^{(t-k)} \stackrel{\text{def}}{=} \boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T$. Then

$$\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})} = \frac{1}{2}\sum_{k=1}^{t\wedge\vartheta}\left[\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k - \mathbb{E}\left[\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k|\mathcal{F}_{k-1}\right]\right]. \tag{101}$$

In view of union bounds, it suffices to impose bounds on each summand of (101) since $k = 1, \cdots, t\wedge\vartheta$ and $t \le T$. We first specify K as the absolute constant in Definition 17. In light of Remark 20 we replace K with $K \cdot M_k$ where $M_k \stackrel{\text{def}}{=} \max_{i\in[n]} |(\boldsymbol{P}_k\boldsymbol{w}_{k-1}^{\vartheta})_i|$. By our choice of α such that $\alpha > \alpha_0 + \theta$ and $\alpha < \alpha'$, then we get from Lemma 9 with α,

$$\max_{1\le k\le T}\max_{1\le i\le n}|\boldsymbol{e}_i^T\boldsymbol{w}_{k-1}^{\vartheta}| \le C(T)n^{\alpha-1/2}$$

and we obtain

$$M_k = \max_{i\in[n]}|(\boldsymbol{X}_k)_i| = \max_{i\in[n]}|(\boldsymbol{P}_k\boldsymbol{w}_{k-1}^{\vartheta})_i| \le \max_{1\le k\le T}\max_{1\le i\le n}|\boldsymbol{e}_i^T\boldsymbol{w}_{k-1}^{\vartheta}| \le C(T)n^{\alpha-1/2}. \tag{102}$$

In order to apply Lemma 19, we need to compute $\|\boldsymbol{D}^{(t-k)}\|_{HS}$ and $\|\boldsymbol{D}^{(t-k)}\|$. Now observe that

$$\|\boldsymbol{D}^{(t-k)}\|_{HS}^2 \le \|\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\|^2\|\boldsymbol{A}\|_{HS}^2 \le \|\boldsymbol{A}\|^4\|\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\|^2\cdot n \le C(T)n \tag{103}$$

and

$$\|\boldsymbol{D}^{(t-k)}\|_2 = \|\boldsymbol{A}\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\boldsymbol{A}^T\| \le \|\boldsymbol{A}\|^2\|\widetilde{\mathbf{N}}_{(t\wedge\vartheta,k)}\| \le C(T). \tag{104}$$

We note that $t \le T$ and $\|\widetilde{Q}^{(t\wedge k)}\|$ can be upper bounded by constants independent of n and d. Lastly, we bound $\|\boldsymbol{D}^{(t-k)}\,\mathbb{E}[\boldsymbol{X}_k|\mathcal{F}_{k-1}]\|$ where $\mathbb{E}[\boldsymbol{X}_k\,|\,\mathcal{F}_{k-1}] = (\mu_1,\ldots,\mu_n)$ and $\mu_\ell = \zeta\boldsymbol{w}_{k-1,\ell}^{\vartheta}$. Using the definition of the stopping time ϑ and (104) we obtain

$$\|\boldsymbol{D}^{(t-k)}\,\mathbb{E}[\boldsymbol{X}_k|\mathcal{F}_{k-1}]\| \le \|\boldsymbol{D}^{(t-k)}\|\cdot\|\,\mathbb{E}[\boldsymbol{X}_k\,|\,\mathcal{F}_{k-1}]\| = \zeta\|\boldsymbol{D}^{(t-k)}\|\|\boldsymbol{w}_{k-1}^{\vartheta}\| \le C(T)n^{\theta} \tag{105}$$

Using Lemma 19 (Hanson-Wright) and the remark following it, we have

$$\begin{aligned}
&\Pr(|\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k - \mathbb{E}\,\boldsymbol{X}_k^T\boldsymbol{D}^{(t-k)}\boldsymbol{X}_k| \ge \tilde{\epsilon}\,|\,\mathcal{F}_{k-1})\\
&\qquad\le 2\exp\left(-\frac{1}{C}\min\left(\frac{\tilde{\epsilon}^2}{2\cdot 9M_k^4K^4\|\boldsymbol{D}^{(t-k)}\|_{HS}^2}, \frac{\tilde{\epsilon}}{3M_k^2K^2\|\boldsymbol{D}^{(t-k)}\|}\right)\right)\\
&\qquad\quad + 2\cdot 2\exp\left(-\frac{\tilde{\epsilon}^2}{9M_k^2K^2\|\boldsymbol{D}^{(t-k)}(\mathbb{E}[\boldsymbol{X}_k\,|\,\mathcal{F}_{k-1}])\|^2}\right).
\end{aligned}$$

Let $\tilde{\epsilon} = n^{2\alpha'-1/2}\cdot 2(T)^{-1}$. By using (102), (103), (104), and (105),

$$\frac{\tilde{\epsilon}^2}{2\cdot 9M_k^4K^4\|\boldsymbol{D}^{(t-k)}\|_{HS}^2} \ge C(T)\cdot\frac{n^{4\alpha'-1}}{n^{4\alpha-2}n} = C(T)\cdot\frac{1}{n^{4(\alpha-\alpha')}}, \tag{106}$$

$$\frac{\tilde{\epsilon}}{3M_k^2K^2\|\boldsymbol{D}^{(t-k)}\|} \ge C(T)\cdot\frac{n^{2\alpha'-1/2}}{n^{2\alpha-1}} = C(T)\cdot\frac{1}{n^{2(\alpha-\alpha')-1/2}}, \tag{107}$$

and

$$\frac{\tilde{\epsilon}^2}{M_k^2K^2\|\boldsymbol{D}^{(t-k)}(\mathbb{E}[\boldsymbol{X}_k\,|\,\mathcal{F}_{k-1}])\|_2^2} \ge C(T)\cdot\frac{n^{4\alpha'-1}}{n^{2\alpha-1}n^{2\theta}} = C(T)\cdot\frac{1}{n^{2\alpha-4\alpha'+2\theta}}, \tag{108}$$

where $C(T)$ is independent of n and k, and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Therefore by our choice of $\alpha' > \alpha$ and $0 < \theta < 2\alpha' - \alpha$, we get

$$\boldsymbol{P}(|\boldsymbol{X}_k^T \boldsymbol{D} \boldsymbol{X}_k - \mathbb{E}\, \boldsymbol{X}_k^T \boldsymbol{D} \boldsymbol{X}_k| \geq \tilde{\epsilon} | \mathcal{F}_{k-1}) \leq 2 \exp\left(-C(T) n^c\right) \qquad \text{when } \tilde{\epsilon} = \frac{2n^{2\alpha'-1/2}}{T} \tag{109}$$

for some $c > 0$ and $T > 0$. We set $\epsilon = n^{2\alpha'-1/2}$. Applying two union bounds and (109), we get

$$\begin{aligned}
\Pr\left(\max_{0\leq t\leq T} |\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})}| \geq \epsilon \right) &\leq \sum_{t=0}^{T} \Pr\left(|\mathcal{E}_{t\wedge\vartheta}^{(\nabla^2:\mathrm{HW})}| \geq n^{2\alpha'-1/2} \right) \\
&\leq \sum_{t=0}^{T} \Pr\left(\frac{1}{2} \sum_{k=1}^{t\wedge\vartheta} \left| \boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k | \mathcal{F}_{k-1}] \right| \geq \epsilon \right) \\
&\leq \sum_{t=0}^{T} \sum_{k=1}^{t\wedge\vartheta} \Pr\left(\left| \boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k - \mathbb{E}[\boldsymbol{X}_k^T \boldsymbol{D}^{(t-k)} \boldsymbol{X}_k | \mathcal{F}_{k-1}] \right| \geq \frac{2n^{2\alpha'-1/2}}{T} \right) \\
&\leq 2T^2 \exp(-C(T) n^c)
\end{aligned} \tag{110}$$

for some $c, C(T) > 0$. In the penultimate inequality, we used that $t \wedge \vartheta \leq T$. The result immediately follows. ■

Now we return back to Proposition 2.2 as we have shown all the components are small.

Proof [Proof of Proposition 2.2] Let $\alpha_0 \in (0, 1/4)$ as specified in Assumption 4. Let α, α', and θ satisfying $0 < \theta < \alpha' - \alpha$ and $\alpha_0 + \theta < \alpha < \alpha'$. Then applying the decomposition in Lemma 11 and the lemmas showing the decomposed error terms are small when n and d are large (i.e. Lemma 13, 14, 16, 21, and Corollary 15) we get

$$\begin{aligned}
\max_{0\leq t\leq T} |\mathcal{E}_t^{\nabla^2\text{-Diag}}| &\leq \max_{0\leq t\leq T} |\mathcal{E}_t^{(\nabla^2:\mathrm{KL})}| + \max_{0\leq t\leq T} |\mathcal{E}_t^{(\nabla^2:\mathrm{B.1})}| + \max_{0\leq t\leq T} |\mathcal{E}_t^{(\nabla^2:\mathrm{B.2})}| \\
&\quad + \max_{0\leq t\leq T} |\mathcal{E}_t^{(\nabla^2:\mathrm{Z.1})}| + \max_{0\leq t\leq T} |\mathcal{E}_t^{(\nabla^2:\mathrm{Z.2})}| + \max_{0\leq t\leq T} |\mathcal{E}_t^{(\nabla^2:\mathrm{HW})}| \\
&\leq C(T) n^{\alpha_0 - 1/2 + 2\theta} + 2C(T) n^{\alpha' - 1/2} + 2C(T) n^{2\theta - 1} + C(T) n^{2\alpha' - 1/2} \\
&\leq C(T) n^{2\alpha' - 1/2}
\end{aligned} \tag{111}$$

holds with overwhelming probability. ■

2.5. Proof of Proposition 2.3

In this section, we control the off-diagonal martingale error term

$$\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}} = \sum_{1\leq k_1<k_2}^{t\wedge\vartheta} \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{N}_{(t\wedge\vartheta,k_1)} (\nabla^2\mathcal{R}) \mathbf{N}_{(t\wedge\vartheta,k_2)} \mathring{\boldsymbol{M}}_{k_2}, \tag{112}$$

where we recall the martingale increment (20) and the matrix $\mathbf{N}_{t,k}$ from Proposition **??**, respectively,

$$\mathring{\boldsymbol{M}}_k = \zeta \boldsymbol{A}^T \boldsymbol{w}_{k-1} - \boldsymbol{A}^T \boldsymbol{P}_k \boldsymbol{w}_{k-1} \quad \text{and} \quad \mathbf{N}_{t,k} = \mathbf{N}_{t,k}(\boldsymbol{H}). \tag{113}$$

We will show that this error term is small, the main result of this section, proof of Proposition 2.3.

To prove Proposition 2.3, we first fix the k_2 terms and show the resulting summation is small using Bernstein concentration (Proposition 2.5). To do so, we need to show certain terms are themselves small, which depend on the martingale increment $\mathring{\boldsymbol{M}}_{k_2}$. This will require us to use Hanson-Wright, Lemma 19. Combining both Proposition 2.5 (Bardenet) and Lemma 19 (Hanson-Wright), Proposition 2.3 will follow

To this end, for convenience, we define

$$\boldsymbol{V}^{(t\wedge\vartheta,k_2)} \stackrel{\text{def}}{=} \sum_{k_1=1}^{k_2-1} \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{N}_{(t\wedge\vartheta,k_1)}(\nabla^2\mathcal{R})\mathbf{N}_{(t\wedge\vartheta,k_2)} \qquad k_2 = 1,2,\cdots t\wedge\vartheta \tag{114}$$

$$\boldsymbol{Y}^{(t\wedge\vartheta,k_2)} \stackrel{\text{def}}{=} \sum_{j\in B_{k_2-1}} \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j (\boldsymbol{A}\boldsymbol{x}_{k_2-1}-\boldsymbol{b})_j \qquad k_2 = 1,2,\cdots t\wedge\vartheta \tag{115}$$

$$\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} \stackrel{\text{def}}{=} \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j (\boldsymbol{A}\boldsymbol{x}_{k_2-1}-\boldsymbol{b})_j \qquad k_2 = 1,2,\cdots t\wedge\vartheta. \tag{116}$$

The following lemma and its corollary will be useful in showing that the off-diagonal term is small.

Lemma 22 *Let $\delta > 0$ such that $\alpha_0 + \theta < \delta < 1/4$ and $0 < \theta < \frac{1}{4} - \delta$. Then the following holds with overwhelming probability*

$$\max_{1\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\|^2 \le C(T)\cdot n^{1-2\delta}. \tag{117}$$

Proof We have

$$\begin{aligned}
&\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\| = \max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \left\| \left(\sum_{k_1=1}^{k_2-1} \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{N}_{(t\wedge\vartheta,k_1)}(\nabla^2\mathcal{R})\mathbf{N}_{(t\wedge\vartheta,k_2)} \right)\boldsymbol{A}^T \right\| \\
&\le T \max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta}\max_{1\le k_1\le k_2-1} \left\| \left(\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{N}_{(t\wedge\vartheta,k_1)} \left(\nabla^2\mathcal{R}\right) \mathbf{N}_{(t\wedge\vartheta,k_2)} \right)\boldsymbol{A}^T \right\| \\
&\le T \max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta}\max_{1\le k_1\le k_2-1} \|\mathring{\boldsymbol{M}}_{k_1}\| \cdot \|\mathbf{N}_{(t\wedge\vartheta,k_1)}\| \cdot \|\mathbf{N}_{(t\wedge\vartheta,k_2)}\| \cdot \|\boldsymbol{A}\| \cdot \|\nabla^2\mathcal{R}\| \\
&\le T \left(\max_{1\le k_1\le (T\wedge\vartheta)-1} \|\mathring{\boldsymbol{M}}_{k_1}\| \right) \cdot \left(\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\mathbf{N}_{(t\wedge\vartheta,k_2)}\|^2 \cdot \|\boldsymbol{A}\| \cdot \|\nabla^2\mathcal{R}\| \right) \\
&\le C(T)\cdot \max_{1\le k_1\le (T\wedge\vartheta)-1} \|\mathring{\boldsymbol{M}}_{k_1}\|
\end{aligned} \tag{118}$$

which implies

$$\max_{0\le t\le T}\max_{1\le k_2\le t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\|^2 \le C(T)\cdot \max_{1\le k_1\le (T\wedge\vartheta)-1} \|\mathring{\boldsymbol{M}}_{k_1}\|^2 \tag{119}$$

where the constant C is dependent on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta, |\Omega|$, and time T, but it is independent of n or d.

Fix $k_1 \le (T\wedge\vartheta)-1$. By adding and subtracting the mean of the quadratic martingale term and applying the triangle inequality we get

$$\|\mathring{\boldsymbol{M}}_{k_1}\|^2 \le \left| \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \mid \mathcal{F}_{k_1-1} \right] \right| + \left| \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \mid \mathcal{F}_{k_1-1} \right] \right|. \tag{120}$$

We address the first term in the sum by applying Lemma 19 (Hanson-Wright). We specify K as the absolute constant in Definition 17. In light of remark 20, we replace K with $K \cdot M_{k_1}$ where $M_{k_1} \overset{\text{def}}{=} \max_{1\le i\le n} |\mathring{\boldsymbol{M}}_{k_1,i}|$. By the definition of the stopping time ϑ and the fact that $k_1 \le (T \wedge \vartheta) - 1$ we get

$$\begin{aligned} M_{k_1} &= \max_{1\le i\le n} |\mathring{\boldsymbol{M}}_{k_1,i}| \\ &= \max_{1\le i\le n} |(\zeta \boldsymbol{A}^T \boldsymbol{w}_{k_1-1})_i - (\boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k_1-1})_i| \\ &\le \max_{1\le i\le n} \zeta |\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k_1-1}| + \max_{1\le i\le n} |\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k_1-1}| \\ &\le \zeta \cdot \|\boldsymbol{A}\| \cdot \|\boldsymbol{w}_{k_1-1}\| + \|\boldsymbol{A}\| \cdot \|\boldsymbol{w}_{k_1-1}\| \\ &\le C(\zeta, \|\boldsymbol{A}\|) \cdot n^{\theta} \end{aligned} \tag{121}$$

where $C(\zeta, \|\boldsymbol{A}\|)$ is independent of n or d. Choosing $\epsilon = n^{1-2\delta}$ and applying Lemma 19 we get

$$\Pr\left(\left| \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k_1-1} \right] \right| \ge \epsilon \right) \le 2 \exp\left(-\frac{1}{D} \min\left(\frac{\epsilon^2}{2M_{k_1}^4 K^4 \|\mathbf{I}\|_{HS}^2}, \frac{\epsilon}{M_{k_1}^2 K^2 \|\mathbf{I}\|} \right) \right). \tag{122}$$

for some universal constant $D > 0$. Moreover we have that

$$\frac{\epsilon^2}{2M_{k_1}^4 K^4 \|\mathbf{I}\|_{HS}^2} \ge C \cdot \frac{n^{2-4\delta}}{n^{4\theta} n} = C \cdot \frac{1}{n^{4(\delta+\theta)-1}}$$

$$\text{and} \qquad \frac{\epsilon}{M_{k_1}^2 K^2 \|\mathbf{I}\|} \ge C \cdot \frac{n^{1-2\delta}}{n^{2\theta}} = C \cdot \frac{1}{n^{2(\delta+\theta)-1}}$$

where $C > 0$ is independent of n and d. By our assumptions on δ and θ, namely $\delta + \theta < 1/4$, we have

$$\Pr\left(\left| \mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k_1-1} \right] \right| \ge n^{1-2\delta} \right) \le \exp(-Cn^c) \tag{123}$$

for some $c > 0$ and constant C independent of n and d.

Now we bound $\mathbb{E}\left[\|\mathring{\boldsymbol{M}}_{k_1}\|^2 \,|\, \mathcal{F}_{k_1-1} \right]$ in (120). From (86), we know that

$$\begin{aligned} \mathring{\boldsymbol{M}}_{k_1,i} \mathring{\boldsymbol{M}}_{k_1,i} &= (\zeta - \zeta^2) \sum_{\ell=1}^{n} A_{\ell i} A_{\ell i} w_{k_1-1,\ell}^2 + \left(\frac{\beta(\beta-1)}{n(n-1)} - \zeta^2 \right) \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k_1-1} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{w}_{k_1-1} \\ &+ \left(\left[\frac{\beta}{n} - \frac{\beta(\beta-1)}{n(n-1)} \right] - (\zeta - \zeta^2) \right) \sum_{\ell=1}^{n} A_{\ell i} A_{\ell i} w_{k_1-1,\ell}^2 + \mathcal{E}_{k_1}^{(\nabla^2,1)}(i,i) + \mathcal{E}_{k_1}^{(\nabla^2,2)}(i,i) \\ &+ \left(\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k_1-1} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k-1} - \mathbb{E}\left[\boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k-1} \boldsymbol{e}_i^T \boldsymbol{A}^T \boldsymbol{P}_{k_1} \boldsymbol{w}_{k_1-1} | \mathcal{F}_{k_1-1} \right] \right), \end{aligned} \tag{124}$$

where $\mathcal{E}_{k_1}^{(\nabla^2,1)}(i,i)$ and $\mathcal{E}_{k_1}^{(\nabla^2,2)}(i,i)$ are defined in (82). When we take conditional expectation, we see that $\mathbb{E}[\mathcal{E}_{k_1}^{(\nabla^2,1)}(i,i)\,|\,\mathcal{F}_{k_1-1}], \mathbb{E}[\mathcal{E}_{k_1}^{(\nabla^2,2)}(i,i)\,|\,\mathcal{F}_{k_1-1}] = 0$. Therefore, we deduce that

$$\begin{aligned}
\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1,i}\mathring{\boldsymbol{M}}_{k_1,i}|\mathcal{F}_{k_1-1}\right] &= (\zeta-\zeta^2)\sum_{\ell=1}^{n} A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 + \left(\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2\right)\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1} \\
&\quad + \left(\left[\frac{\beta}{n}-\frac{\beta(\beta-1)}{n(n-1)}\right]-(\zeta-\zeta^2)\right)\sum_{\ell=1}^{n} A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 \\
&= (\zeta-\zeta^2)\sum_{\ell=1}^{n} A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 + \frac{\zeta(\zeta-1)}{n-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1} \\
&\quad - \frac{\zeta(\zeta-1)}{n-1}\sum_{\ell=1}^{n} A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2.
\end{aligned} \tag{125}$$

where in the last line we used the fact $\frac{\beta(\beta-1)}{n(n-1)}-\zeta^2 = \frac{\zeta(\zeta-1)}{n-1}$. By summing over i and applying the definition of the stopping time, Lemma 10, the fact that $\zeta\in(0,1)$, and that $k_1 \le (T\wedge\vartheta)-1$ we get

$$\begin{aligned}
\left|\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{I}\mathring{\boldsymbol{M}}_{k_1}\,|\,\mathcal{F}_{k_1-1}\right]\right| &\le \left|\frac{\zeta(\zeta-1)}{n-1}\right|\sum_{i=1}^{d}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\boldsymbol{e}_i^T\boldsymbol{A}^T\boldsymbol{w}_{k_1-1} + \left|\frac{\zeta(\zeta-1)}{n-1}\right|\sum_{i=1}^{d}\sum_{\ell=1}^{n}A_{\ell i}A_{\ell i}w_{k_1-1,\ell}^2 \\
&\quad + (\zeta-\zeta^2)\sum_{i=1}^{d}\sum_{\ell=1}^{n}\boldsymbol{A}_{\ell i}^2(\boldsymbol{w}_{k_1-1,\ell})^2 \\
&= \left|\frac{\zeta(\zeta-1)}{n-1}\right|\sum_{i=1}^{d}\left(\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\right)_i^2 + \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right|+(\zeta-\zeta^2)\right)\sum_{\ell=1}^{n}\boldsymbol{w}_{k_1-1,\ell}^2\left(\sum_{i=1}^{d}\boldsymbol{A}_{\ell i}^2\right) \\
&= \left|\frac{\zeta(\zeta-1)}{n-1}\right|\|\boldsymbol{A}^T\boldsymbol{w}_{k_1-1}\|^2 + \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right|+(\zeta-\zeta^2)\right)\sum_{\ell=1}^{n}\boldsymbol{w}_{k_1-1,\ell}^2\left(\boldsymbol{A}\boldsymbol{A}^T\right)_{\ell\ell} \\
&\le \left|\frac{\zeta(\zeta-1)}{n-1}\right|\|\boldsymbol{A}\|^2\|\boldsymbol{w}_{k_1-1}\|^2 + \left(\left|\frac{\zeta(\zeta-1)}{n-1}\right|+(\zeta-\zeta^2)\right)\max_{1\le j\le n}\left(\boldsymbol{A}\boldsymbol{A}^T\right)_{jj}\|\boldsymbol{w}_{k_1-1}\|^2 \\
&\le C(T)n^{2\theta-1} \\
&\quad + C(T)n^{2\theta}\left(\left|\frac{\zeta(\zeta-1)}{n-1}\right|+(\zeta-\zeta^2)\right)\left|\max_{j\in[n]}\left(\boldsymbol{A}\boldsymbol{A}^T\right)_{jj}+\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)-\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right| \\
&\le C(T)n^{2\theta-1} \\
&\quad + C(T)n^{2\theta}\left(\left|\frac{\zeta(\zeta-1)}{n-1}\right|+(\zeta-\zeta^2)\right)\left(\left|\max_{j\in[n]}\left(\boldsymbol{A}\boldsymbol{A}^T\right)_{jj}-\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right|+\left|\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right|\right) \\
&\le C(T)n^{2\theta-1} + C(T)n^{2\theta}\left(\left|\frac{\zeta(\zeta-1)}{n-1}\right|+(\zeta-\zeta^2)\right)\left(C(T)n^{\alpha_0-1/2}+\left|\frac{1}{n}\operatorname{tr}\left(\boldsymbol{A}\boldsymbol{A}^T\right)\right|\right) \\
&\le C(T)n^{2\theta-1} + C(T)n^{\alpha_0+2\theta-1/2} + C(T)n^{2\theta}.
\end{aligned} \tag{126}$$

In the last inequality, we used that $\frac{1}{n}\text{tr}(\boldsymbol{A}\boldsymbol{A}^T)$ is independent of n and d by our assumptions on the data matrix $\boldsymbol{A}$. By the assumptions on θ and δ, we have that

$$\left|\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k_1-1}\right]\right| \leq C(T) n^{1-2\delta}, \tag{127}$$

where $C(T)$ is independent of n and d, and only depends on $\|\boldsymbol{A}\|, \|\boldsymbol{b}\|, \|\boldsymbol{x}_0\|, |\Omega|, \|\mathcal{R}\|_{H^2}, \gamma, \zeta$ and T. Putting everything together, we have from (118), (123), and (127) that

$$\begin{aligned}
\max_{0\leq t\leq T} \max_{1\leq k_2\leq t\wedge\vartheta} \|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\| &\leq C(T) \max_{1\leq k_1\leq (T\wedge\vartheta)-1} \|\mathring{\boldsymbol{M}}_{k_1}\| \\
&\leq C(T)\left(\max_{1\leq k_1\leq t\wedge\vartheta} \left|\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} - \mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k-1}\right]\right| + \max_{1\leq k_1\leq t\wedge\vartheta} \left|\mathbb{E}\left[\mathring{\boldsymbol{M}}_{k_1}^T \mathbf{I} \mathring{\boldsymbol{M}}_{k_1} \,|\, \mathcal{F}_{k-1}\right]\right| \right) \\
&\leq C(T) n^{1-2\delta} \qquad \text{w.o.p.}
\end{aligned} \tag{128}$$

The result is now shown. ■

We now show that $X_j^{(t\wedge\vartheta,k_2)}$ (116) is small.

Lemma 23 *For all α and δ such that $\alpha_0 + \theta < \alpha < \delta < 1/4$ and $0 < \theta < \frac{1}{4} - \delta$ we have*

$$\max_{0\leq t\leq T} \frac{1}{n}\sum_{j=1}^n |\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2 \leq C(T) n^{2(\alpha-\delta)-1} \quad \textit{and} \quad \max_{0\leq t\leq T} \max_{j\in[n]} |\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}| \leq C(T) n^{\alpha-\delta}, \tag{129}$$

for $0 \leq k_2 \leq t\wedge\vartheta$.

Proof Let $0 \leq t \leq T$ and $0 \leq k_2 \leq t\wedge\vartheta$. We achieve the first result in (129) as follows

$$\begin{aligned}
\sum_{j=1}^n |\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2 &= \sum_{j=1}^n \left| \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j (\boldsymbol{A}\boldsymbol{x}_{k_2-1} - \boldsymbol{b})_j \right|^2 \\
&\leq \max_{1\leq j\leq n} |\boldsymbol{w}_{k_2,j}^{\vartheta}|^2 \sum_{j=1}^n \left| \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j \right|^2 = \max_j |\boldsymbol{w}_{k_2,j}^{\vartheta}|^2 \cdot ||\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T||^2 \\
&\leq C(t) \cdot n^{2\alpha-1} \cdot n^{1-2\delta} = C(t) n^{2(\alpha-\delta)}
\end{aligned} \tag{130}$$

where we applied Lemma 9 and Lemma 22 in the penultimate step. This immediately gives us

$$\max_{0\leq t\leq T} \frac{1}{n}\sum_{j=1}^n |\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2 \leq C(T) \cdot n^{2(\alpha-\delta)-1} \tag{131}$$

where $C(T) = \max_{0\leq t\leq T} C(t)$. This gives the first inequality in (129).

For the second inequality in (129), using (130), we get

$$\max_{1\leq j\leq n} |\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}| \leq \sqrt{\sum_{j=1}^n |\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|^2} \leq C(t) n^{\alpha-\delta}$$

which immediately yields

$$\max_{0\leq t\leq T}\max_{1\leq j\leq n}|\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}|\leq C(T)n^{\alpha-\delta} \tag{132}$$

where again $C(t)$ is independent of n and d and $C(T)=\max_{0\leq t\leq T}C(t)$ and the result is shown. ■

Now we have all the results in order to prove the main proposition, Proposition 2.3, of this section.

Proof [Proof of Proposition 2.3] Let $0\leq t\leq T$. First, we rewrite $\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}$ so that we can apply Lemma 23. To this end, we have

$$\begin{aligned}\mathcal{E}_{t\wedge\vartheta}^{\nabla^2\text{-Off}}&=\sum_{k_1<k_2}^{t\wedge\vartheta}\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{N}_{(t\wedge\vartheta,k_1)}(\nabla^2R)\mathbf{N}_{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}\\&=\sum_{k_2=1}^{t\wedge\vartheta}\left(\sum_{k_1=1}^{k_2-1}\mathring{\boldsymbol{M}}_{k_1}^T\mathbf{N}_{(t\wedge\vartheta,k_1)}\left(\nabla^2\mathcal{R}\right)\mathbf{N}_{(t\wedge\vartheta,k_2)}\right)\mathring{\boldsymbol{M}}_{k_2}\\&=\sum_{k_2=1}^{t\wedge\vartheta}\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2},\end{aligned} \tag{133}$$

where $\boldsymbol{V}^{(t\wedge\vartheta,k_2)}$ is defined in (114). For each $1\leq k_2\leq t\wedge\vartheta$, we get

$$\begin{aligned}|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|&=\left|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\boldsymbol{P}_{k_2}\boldsymbol{w}_{k_2-1}^{\vartheta}-\mathbb{E}\left[\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\boldsymbol{P}_{k_2}\boldsymbol{w}_{k_2-1}^{\vartheta}|\mathcal{F}_{k_2-1}\right]\right|\\&=\left|\sum_{j\in B_{k_2-1}}\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\boldsymbol{w}_{k_2-1,j}^{\vartheta}-\mathbb{E}\left[\sum_{j\in B_{k_2-1}}\left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j\boldsymbol{w}_{k_2-1,j}^{\vartheta}|\mathcal{F}_{k_2-1}\right]\right|\\&=\left|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}-\mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}|\mathcal{F}_{k_2-1}\right]\right|,\end{aligned} \tag{134}$$

where $\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}$ is defined in (115). Fix $\tilde{\epsilon}>0$. Using (134), we have

$$\begin{aligned}\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_t|>\tilde{\epsilon}\right)&=\mathbb{E}\left[\mathbf{1}_{|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|>\tilde{\epsilon}}\right]\\&=\mathbb{E}\left[\mathbb{E}\left[\mathbf{1}_{|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|>\tilde{\epsilon}}\,|\,\mathcal{F}_{k_2-1}\right]\right]\\&=\mathbb{E}\left[\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}|>\tilde{\epsilon}\,|\,\mathcal{F}_{k_2-1}\right)\right]\\&=\mathbb{E}\left[\Pr\left(|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}-\mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}\,|\,\mathcal{F}_{k_2-1}\right]|>\tilde{\epsilon}\right)\right].\end{aligned} \tag{135}$$

This means we can work with the quantity $\Pr\left(|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}-\mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}\,|\,\mathcal{F}_{k_2-1}\right]|>\tilde{\epsilon}\right)$ which allows us to apply Proposition 2.5 [5]. In light of the syntax of Proposition 2.5 and for $1\leq k_2\leq$

$t \wedge \vartheta$, we use $\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} = \left(\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\boldsymbol{A}^T\right)_j (\boldsymbol{A}\boldsymbol{x}_{k_2-1} - \boldsymbol{b})_j$ defined in (116), and define

$$b_{t\wedge\vartheta,k_2} \stackrel{\text{def}}{=} \max_{1\le j\le n} \boldsymbol{X}_j^{(t\wedge\vartheta,k_2)}, \quad \mu_{(t\wedge\vartheta,k_2)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n} \boldsymbol{X}_j^{(t\wedge\vartheta,k_2)},$$

$$\text{and} \qquad \sigma^2_{(t\wedge\vartheta,k_2)} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{j=1}^{n}\left(\boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} - \mu_{(t\wedge\vartheta,k_2)}\right)^2.$$

By Lemma 23, we have the following upper bounds that hold with overwhelming probability

$$\left|b_{(t\wedge\vartheta,k_2)}\right| \le C(T)\cdot n^{\alpha-\delta} \qquad \text{and} \qquad \sigma^2_{(t\wedge\vartheta,k_2)} \le C(T)\cdot n^{2(\alpha-\delta)-1} \tag{136}$$

for some constant $C(T)$ depending on $||\Omega||$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, and T but independent of n and d. Applying Proposition 2.5 yields

$$\begin{aligned}
\Pr\left(\left|\boldsymbol{Y}^{(t\wedge\vartheta,k_2)} - \mathbb{E}\left[\boldsymbol{Y}^{(t\wedge\vartheta,k_2)}|\mathcal{F}_{k_2-1}\right]\right| > \tilde{\epsilon}\right) &= \Pr\left(\left|\sum_{j\in B_{k_2-1}} \boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} - \beta\mu_{t\wedge\vartheta,k_2}\right| \ge \tilde{\epsilon}\right) \\
&= \Pr\left(\left|\frac{1}{\beta}\sum_{j\in B_{k_2-1}} \boldsymbol{X}_j^{(t\wedge\vartheta,k_2)} - \mu_{t\wedge\vartheta,k_2}\right| \ge \tilde{\epsilon}/\beta\right) \\
&\le 2\exp\left(-\frac{\beta(\frac{\tilde{\epsilon}}{\beta})^2}{2\sigma^2_{(t\wedge\vartheta,k_2)} + \frac{2}{3}(b_{(t\wedge\vartheta,k_2)} - a)(\frac{\tilde{\epsilon}}{\beta})}\right)
\end{aligned} \tag{137}$$

Let $\tilde{\epsilon} := n^{\alpha-\delta+\eta}\cdot T^{-1}$ Using (136), we obtain

$$\frac{\beta(\frac{\tilde{\epsilon}}{\beta})^2}{2\sigma^2_{(t\wedge\vartheta,k_2)} + \frac{2}{3}(b_{(t\wedge\vartheta,k_2)} - a)(\frac{\tilde{\epsilon}}{\beta})} \ge C(T)\frac{T^{-2}\cdot n^{-1}n^{2(\alpha-\delta+\eta)}}{n^{2(\alpha-\delta)-1} + T^{-1}\cdot n^{\alpha-\delta}n^{-1}n^{\alpha-\delta+\eta}} \ge C(T)\frac{1}{n^{-2\eta}+n^{-\eta}} \to \infty.$$

Therefore it follows that

$$\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}| > \tilde{\epsilon}\right) \le 2\exp(-C(T)n^c) \qquad \text{when } \tilde{\epsilon} = T^{-1}\cdot n^{\alpha-\delta+\eta} \tag{138}$$

for some constant $c > 0$ and $C(T)$ where the constants are independent of n and d and only depend on $||\Omega||$, $\|\boldsymbol{A}^T\boldsymbol{A}\|$, γ and T. Here again $C(T)$ is a positive constant independent of n and d. Let $\epsilon = n^{\alpha-\delta+\eta}$. Using (133) and (138), and observing that $\boldsymbol{V}^{(t\wedge\vartheta,k_2)} = \boldsymbol{0}$ for $k_2 > t\wedge\vartheta$ we get

$$\begin{aligned}
\Pr\left(\max_{0\le t\le T}\left|\mathcal{E}^{\nabla^2\text{-Off}}_{t\wedge\vartheta}\right| \ge \epsilon\right) &= \sum_{t=0}^{T}\Pr\left(\sum_{k_2=1}^{t\wedge\vartheta}|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}| \ge \epsilon\right) \\
&\le \sum_{t=0}^{T}\sum_{k_2=1}^{T}\Pr\left(|\boldsymbol{V}^{(t\wedge\vartheta,k_2)}\mathring{\boldsymbol{M}}_{k_2}| \ge \frac{\epsilon}{T}\right) \\
&\le 2\sum_{t=0}^{T}\sum_{k_2=1}^{T}\exp\left(-C(T)n^c\right) = 2T^2\cdot\exp\left(-C(T)n^c\right).
\end{aligned} \tag{139}$$

This gets used our desired result. ■