Subquantile Minimization for Kernel Learning in the Huber ϵ-Contamination Model

Arvind Rathnashyam* Alex Gittens†

October 1, 2023

Abstract

In this paper we propose Subquantile Minimization for learning with adversarial corruption in the training set. Superquantile objectives have been formed in the past in the context of fairness where one wants to learn an underrepresented distribution equally [17, 31]. Our intuition is to learn a more favorable representation of the *majority* class, thus we propose to optimize over the p-subquantile of the loss in the dataset. In particular, we study the Huber Contamination Problem for Kernel Learning where the distribution is formed as, $\hat{\mathbb{P}} = (1-\varepsilon)\mathbb{P} + \varepsilon\mathbb{Q}$, and we want to find the function $\inf_f \mathbb{E}_{\mathbf{x}\in\mathbb{P}}[\ell_f(\mathbf{x})]$, from the noisy distribution, $\hat{\mathbb{P}}$. We assume the adversary has knowledge of the true distribution of $\mathbb{P}$, and is able to corrupt the covariates and the labels of ε samples. To our knowledge, we are the first to study the problem of general kernel learning in the Huber Contamination Model. In our theoretical analysis, we analyze our non-convex concave objective function with the Moreau Envelope. We show (i) a stationary point with respect to the Moreau Envelope is a good point and (ii) we can reach a stationary point with gradient descent methods. Further, we analyze accelerated gradient methods for the non-convex concave minimax optimization problem. We empirically test Kernel Ridge Regression and Kernel Classification on various state of the art datasets and show Subquantile Minimization gives strong results. Furthermore, we run experiments on various datasets and compare with the state-of-the-art algorithms to show the superior performance of Subquantile Minimization.

*CS, Rensselear Polytechnic Institute, rathna@rpi.edu
†CS, Rensselaer Polytechnic Institute, gittea@rpi.edu

References

[1] Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[2] Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[3] Yu Cheng, Ilias Diakonikolas, Rong Ge, and Mahdi Soltanolkotabi. High-dimensional robust mean estimation via gradient descent. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1768–1778. PMLR, 13–18 Jul 2020.

[4] Yu Cheng, Ilias Diakonikolas, Rong Ge, and David P Woodruff. Faster algorithms for high-dimensional robust covariance estimation. In *Conference on Learning Theory*, pages 727–757. PMLR, 2019.

[5] Ilias Diakonikolas, Gautam Kamath, Daniel M. Kane, Jerry Li, Jacob Steinhardt, and Alistair Stewart. Sever: A robust meta-algorithm for stochastic optimization. In *Proceedings of the 36th International Conference on Machine Learning*, ICML '19, pages 1596–1606. JMLR, Inc., 2019.

[6] Ilias Diakonikolas and Daniel M Kane. *Algorithmic high-dimensional robust statistics*. Cambridge University Press, 2023.

[7] Dheeru Dua and Casey" Graff. UCI machine learning repository, 2017.

[8] Jianqing Fan, Weichen Wang, and Yiqiao Zhong. An l∞ eigenvector perturbation bound and its application to robust covariance estimation. *Journal of Machine Learning Research*, 18(207):1–42, 2018.

[9] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, jun 1981.

[10] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[11] Arthur Gretton. Introduction to rkhs, and some simple kernel algorithms. *Adv. Top. Mach. Learn. Lecture Conducted from University College London*, 16(5-3):2, 2013.

[12] Peter J. Huber and Elvezio. Ronchetti. *Robust statistics*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 2nd ed. edition, 2009.

[13] Steinbrunn William Pfisterer Matthias Janosi, Andras and Robert Detrano. Heart Disease. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C52P4X.

[14] Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *ICML*, 2018.

[15] Chi Jin, Praneeth Netrapalli, and Michael Jordan. What is local optimality in nonconvex-nonconcave minimax optimization? In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4880–4889. PMLR, 13–18 Jul 2020.

[16] Ashish Khetan, Zachary C. Lipton, and Anima Anandkumar. Learning from noisy singly-labeled data. In *International Conference on Learning Representations*, 2018.

[17] Yassine Laguel, Krishna Pillutla, Jérôme Malick, and Zaid Harchaoui. Superquantiles at work: Machine learning applications and efficient subgradient computation. *Set-Valued and Variational Analysis*, 29(4):967–996, Dec 2021.

[18] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations*, 2021.

[19] Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France*, 93:273–299, 1965.

[20] Bhaskar Mukhoty, Govind Gopakumar, Prateek Jain, and Purushottam Kar. Globally-convergent iteratively reweighted least squares for robust regression problems. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 313–322. PMLR, 16–18 Apr 2019.

[21] Yurii Evgen'evich Nesterov. A method of solving a convex programming problem with convergence rate o\bigl(k^2\bigr). In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences, 1983.

[22] Muhammad Osama, Dave Zachariah, and Petre Stoica. Robust risk minimization for statistical learning from corrupted data. *IEEE Open Journal of Signal Processing*, 1:287–294, 2020.

[23] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17, 1964.

[24] Adarsh Prasad, Sivaraman Balakrishnan, and Pradeep Ravikumar. A unified approach to robust mean estimation. *arXiv preprint arXiv:1907.00927*, 2019.

[25] Adarsh Prasad, Arun Sai Suggala, Sivaraman Balakrishnan, and Pradeep Ravikumar. Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*, 2018.

[26] Ning Qian. On the momentum term in gradient descent learning algorithms. *Neural networks*, 12(1):145–151, 1999.

[27] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[28] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Processing Magazine*, 37(5):55–66, 2020.

[29] R Tyrrell Rockafellar. Convex analaysis. 2015.

[30] Ralph Tyrell Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

[31] R.T. Rockafellar, J.O. Royset, and S.I. Miranda. Superquantile regression with applications to buffered reliability, uncertainty quantification, and conditional value-at-risk. *European Journal of Operational Research*, 234(1):140–154, 2014.

[32] Jacob Steinhardt, Moses Charikar, and Gregory Valiant. Resilience: A criterion for learning in the presence of arbitrary outliers. In Anna R. Karlin, editor, *9th Innovations in Theoretical Computer Science Conference, ITCS 2018, January 11-14, 2018, Cambridge, MA, USA*, volume 94 of *LIPIcs*, pages 45:1–45:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[33] Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

[34] Roman Vershynin. High-dimensional probability. *University of California, Irvine*, 2020.

[35] Junchi Yang, Antonio Orvieto, Aurelien Lucchi, and Niao He. Faster single-loop algorithms for minimax optimization without strong concavity. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 5485–5517. PMLR, 28–30 Mar 2022.

[36] Rahul Yedida, Snehanshu Saha, and Tejas Prashanth. Lipschitzlr: Using theoretically computed adaptive learning rates for fast convergence. *Applied Intelligence*, 51:1460–1478, 2021.

A Concentration Inequalities

In this section we will give various concentration inequalities on the inlier data. We first restate our assumptions.

Assumption 26. (Sub-Gaussian Design of Covariates). We assume each covariate is drawn i.i.d from a zero-mean covariance $\mathbf{\Sigma}$ sub-Gaussian distribution with sub-Gaussian norm $K_1 \in \mathbb{R}_{++}$.

$$\mathbb{E}\left[\mathbf{x}_i\right] = \mathbf{0} \tag{46}$$

$$\mathbb{E}\left[\mathbf{x}_i\mathbf{x}_i^\top\right] = \mathbf{\Sigma} \tag{47}$$

For all $i \in [n]$,

$$\mathbb{P}\left\{\mathbf{v}^\top\mathbf{x}_i \geq t\right\} \leq \exp\left(-\frac{t^2}{K_1^2}\right) \tag{48}$$

Is this equivalent to

$$\mathbb{P}\left\{\mathbf{v}^\top\mathbf{x}_i \geq t\right\} \leq \exp\left(-\frac{t^2}{K_1^2}\right) \tag{49}$$

Assumption 27. (Sub-Gaussian Design of Optimal Residuals). Recall the residual is defined as $\eta_i \triangleq \mathsf{f}^*_{\mathbf{w}}(\mathbf{x}_i) - y_i$. Then we assume for some $K_2 \in \mathbb{R}_{++}$

$$\mathbb{E}\left[\eta_i\right] = 0 \tag{50}$$

$$\mathbb{P}\left\{|\eta_i| \geq t\right\} \leq 2\exp\left(-\frac{t^2}{K_2^2}\right) \tag{51}$$

A.1 Linear Kernel

We will first begin with the Linear Kernel Case.

Assumption 28. Let the data have dimension d such that Assumption 26 holds. Then

$$n > \frac{1}{18}\left(\frac{d^2K^2}{dK-1}\right)\log(2d) \tag{52}$$

Lemma 29. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have sub-Gaussian design described in Assumption 26 and $\eta_i, \ldots, \eta_n$ have sub-Gaussian design described in Assumption 27. It then follows*

$$\mathbb{E}\left|\sum_{i=1}^n \eta_i \left\|\mathbf{x}_i\right\|_2\right| \leq \sqrt{\frac{nd}{K_1}}\left(\exp\left(\frac{C_1}{nd}\right) + \exp\left(C_2\right)\right) = \mathcal{O}\left(\sqrt{nd}\right) \tag{53}$$

Proof. Note $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^n x_i^2}$ is a $\chi(n)$ random variable, thus $\eta_i \left\|\mathbf{x}_i\right\|$ is symmetric around 0. Then for any $t \geq 0$,

$$\mathbb{P}\left\{\sum_{i=1}^n \eta_i \left\|\mathbf{x}_i\right\| \geq t\right\} = \mathbb{P}\left\{\exp\left(\lambda\sum_{i=1}^n \eta_i \left\|\mathbf{x}_i\right\|\right) \geq \exp\left(\lambda t\right)\right\} \tag{54}$$

$$\overset{\text{Markov}}{\leq} e^{-\lambda t}\mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^n \eta_i \left\|\mathbf{x}_i\right\|\right)\right] \tag{55}$$

$$= e^{-\lambda t}\prod_{i=1}^n \underbrace{\mathbb{E}\left[\exp\left(\lambda\eta_i \left\|\mathbf{x}_i\right\|\right)\right]}_{\Upsilon} \tag{56}$$

Where $\lambda > 0$ and will be chosen later. Now we will upper bound Υ.

$$\Upsilon \triangleq \mathbb{E}\left[\lambda \exp\left(\eta_i \|\mathbf{x}_i\|\right)\right] \leq \mathbb{E}\left[\exp\left(\lambda\left(\frac{\eta_i^2}{2} + \frac{\|\mathbf{x}_i\|^2}{2}\right)\right)\right] \tag{57}$$

$$= \mathbb{E}\left[\exp\left(\lambda\frac{\eta_i^2}{2}\right)\exp\left(\lambda\frac{\|\mathbf{x}_i\|^2}{2}\right)\right] \tag{58}$$

$$\stackrel{\text{Young}}{\leq} \frac{1}{2}\mathbb{E}\left[\exp\left(\lambda\eta_i^2\right) + \exp\left(\lambda\|\mathbf{x}_i\|^2\right)\right] \tag{59}$$

$$= \frac{1}{2}\mathbb{E}\left[\exp\left(\lambda\eta_i^2\right) + \prod_{j=1}^{d}\exp\left(\lambda x_j^2\right)\right] \tag{60}$$

$$= \frac{1}{2}\left(\mathbb{E}\left[\exp\left(\lambda\eta_i^2\right)\right] + \prod_{j=1}^{d}\mathbb{E}\left[\exp\left(\lambda x_j^2\right)\right]\right) \tag{61}$$

$$\stackrel{\zeta_1}{\leq} \frac{1}{2}\exp\left(-\sqrt{\frac{K_1}{nd}}t\right)\left(\exp\left(\frac{C_1}{d}\right) + \exp(C_2)\right) \tag{62}$$

In ζ_1, note η_i is sub-exponential and x_i^2 for all $i \in [n]$ are sub-exponential variables as x_i are sub-Gaussian as in Assumption 26. If we plug this back into Equation (56).

$$\mathbb{P}\left\{\sum_{i=1}^{n}\eta_i\|\mathbf{x}_i\| \geq t\right\} \leq e^{-\lambda t}2^{-n}\left(\exp\left(C_1\lambda^2K_2^2\right) + \exp\left(C_2 d\lambda^2 K_1^2\right)\right)^n \tag{63}$$

$$\leq e^{-\lambda t}2^{-n}\left(2^{n-1}\exp\left(C_1 n\lambda^2K_2^2\right) + 2^{n-1}\exp\left(C_2 nd\lambda^2K_1^2\right)\right) \tag{64}$$

$$= \frac{e^{-\lambda t}}{2}\left(\exp\left(C_1 n\lambda^2K_2^2\right) + \exp\left(C_2 nd\lambda^2K_1^2\right)\right) \tag{65}$$

Here we can choose $\lambda \triangleq \sqrt{\frac{K_1}{nd}}$. If we assume $K_1 = K_2$, we then have

$$\mathbb{P}\left\{\sum_{i=1}^{n}\eta_i\|\mathbf{x}_i\| \geq t\right\} = \frac{1}{2}\exp\left(-\sqrt{\frac{K_1}{nd}}t\right)\left(\exp\left(\frac{C_1}{d}\right) + \exp(C_2)\right) \tag{66}$$

We can now upper bound the expectation

$$\mathbb{E}\left|\sum_{i=1}^{n}\eta_i\|\mathbf{x}_i\|_2\right| = \frac{1}{2}\int_0^\infty \exp\left(-\sqrt{\frac{K_1}{nd}}t\right)\left(\exp\left(\frac{C_1}{nd}\right) + \exp(C_2)\right) \tag{67}$$

$$= \sqrt{\frac{nd}{K_1}}\left(\exp\left(\frac{C_1}{nd}\right) + \exp\left(C_2\right)\right) \tag{68}$$

Upper bound Υ also follows from techniques in [34]. ■

Lemma 30. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ have sub-Gaussian design described in Assumption 26 such that $n > d\log(d)$. It then follows*

$$\mathbb{E}\left[\sigma_{\min}\left(\sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top\right)\right] \geq \sigma_{\min}\left(\mathbf{\Sigma}\right)\left(n - \left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right)\right) = \Omega\left(\sigma_{\min}\left(\mathbf{\Sigma}\right)n\right) \tag{69}$$

and with probability $(1-\delta)$

$$\sigma_{\min}\left(\sum_{k\in[n]}\mathbf{x}_k\mathbf{x}_k^\top\right) \geq \frac{1}{1-\delta}\sigma_{\min}\left(\mathbf{\Sigma}\right)\left(n - \left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right)\right) \tag{70}$$

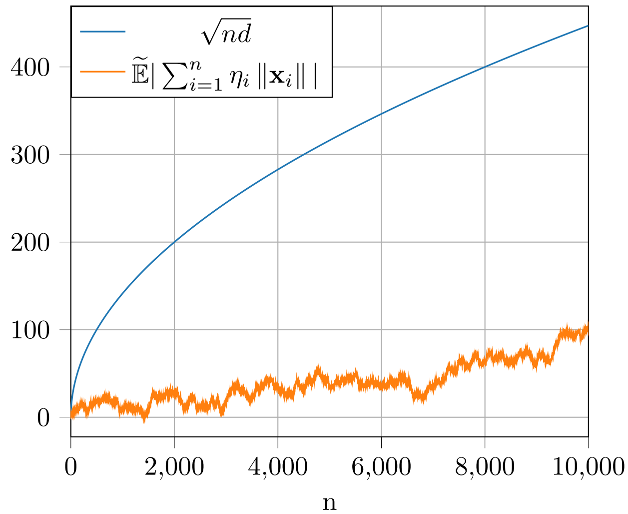

Figure 3: The bound for Lemma 29 compared to an average over 10 trials. We set $\mathbf{\Sigma} = \mathbf{I}$, $K = 4$, $d = 20$.

Proof. In expectation we have

$$\mu_{\min} \triangleq \sigma_{\min}\left(\sum_{i=1}^{n} \mathbb{E}\left[\mathbf{x}_i \mathbf{x}_i^\top\right]\right) = \sigma_{\min}\left(\sum_{i=1}^{n} \mathbf{\Sigma}\right) = n\sigma_{\min}(\mathbf{\Sigma}) \tag{71}$$

Then we have by the sub-Gaussian design in Assumption 26.

$$\sigma_{\min}\left(\sum_{k\in[n]} \mathbf{x}_k \mathbf{x}_k^\top\right) = \sigma_{\min}\left(n\mathbf{\Sigma} + \sum_{k\in[n]} \mathbf{x}_k \mathbf{x}_k^\top - n\mathbf{\Sigma}\right) \tag{72}$$

$$\overset{\text{Weyl's}}{\geq} \sigma_{\min}(n\mathbf{\Sigma}) - \sigma_{\max}\left(\sum_{k\in[n]} \mathbf{x}_k \mathbf{x}_k^\top - n\mathbf{\Sigma}\right) \tag{73}$$

$$= \mu_{\min} - \underbrace{\left\| \sum_{k\in[n]} \mathbf{x}_k \mathbf{x}_k^\top - \mathbf{\Sigma} \right\|_2}_{\text{A}} \tag{74}$$

We assume $\mathbf{\Sigma} \succcurlyeq \mathbf{0}$ and symmetric, therefore we can represent $\mathbf{\Sigma} = (\mathbf{\Sigma}^{1/2})^\top \mathbf{\Sigma}^{1/2} = \mathbf{\Sigma}^{1/2}(\mathbf{\Sigma}^{1/2})^\top$. Next we will upper bound A.

$$\text{A} \triangleq \left\| \sum_{k\in[n]} \mathbf{x}_k \mathbf{x}_k^\top - \mathbf{\Sigma} \right\|_2 \tag{75}$$

$$= \left\| \sum_{k\in[n]} \left(\mathbf{\Sigma}^{1/2}\widehat{\mathbf{x}}_k\right)\left(\mathbf{\Sigma}^{1/2}\widehat{\mathbf{x}}_k\right)^\top - \mathbf{\Sigma} \right\|_2 \tag{76}$$

$$= \left\| \sum_{k\in[n]} \mathbf{\Sigma}^{1/2}\widehat{\mathbf{x}}_k \widehat{\mathbf{x}}_k^\top \left(\mathbf{\Sigma}^{1/2}\right)^\top - \mathbf{\Sigma} \right\|_2 \tag{77}$$

$$= \left\| \mathbf{\Sigma}^{1/2} \left(\sum_{k\in[n]} \mathbf{x}_k \mathbf{x}_k^\top - \mathbf{I}\right) \left(\mathbf{\Sigma}^{1/2}\right)^\top \right\|_2 \tag{78}$$

$$\leq \underbrace{\left\| \sum_{k\in[n]} \widehat{\mathbf{x}}_k \widehat{\mathbf{x}}_k^\top - \mathbf{I} \right\|}_{\text{B}} \|\mathbf{\Sigma}\| \tag{79}$$

It then suffices to upper bound B. First, let $\widehat{\mathbf{x}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and define $K \geq 1$ s.t. $\widehat{\mathbf{x}} \leq dK$, then

$$\begin{aligned}
\mathbb{V}\left(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathbf{I}\right) &= \mathbb{E}\left[\left(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathbf{I}\right)^\top \left(\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top - \mathbf{I}\right)\right] && (80)\\
&= \mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - 2\mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] + \mathbf{I} && (81)\\
&= \mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - \mathbf{I} && (82)\\
&= \mathbb{E}\left[\|\widehat{\mathbf{x}}\|^2 \widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - \mathbf{I} && (83)\\
&\leq dK\mathbb{E}\left[\widehat{\mathbf{x}}\widehat{\mathbf{x}}^\top\right] - \mathbf{I} && (84)\\
&= (dK-1)\,\mathbf{I} && (85)
\end{aligned}$$

Then from Matrix Bernstein, we have

$$\mathbb{E}\left[\text{B}\right] \leq \sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d) \tag{86}$$

Then we have in expectation

$$\mathbb{E}\left[\sigma_{\min}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top\right)\right] \geq n\sigma_{min}\left(\mathbf{\Sigma}\right) - \|\mathbf{\Sigma}\|\left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right) \tag{87}$$

Assume the spectrum of $\mathbf{\Sigma}$ is flat, then we have

$$\mathbb{E}\left[\sigma_{\min}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top\right)\right] \geq \sigma_{\min}\left(\mathbf{\Sigma}\right)\left(n - \left(\sqrt{2(dK-1)n\log(2d)} + \frac{1}{3}dK\log(2d)\right)\right) \tag{88}$$

Similarly from the Matrix Bernstein Inequality in [33], we have

$$\mathbb{P}\left\{\sigma_{\min}\left(\sum_{k\in[n]} \mathbf{x}_k\mathbf{x}_k^\top\right) > t\right\} \leq (n+d)\exp\left(\frac{-t^2}{2n(dK-1) + Kt/3}\right) \tag{89}$$

■

Lemma 31. *Let $\eta_i \in P \cap S$ be defined as in Assumption 27, then it follows*

$$\mathbb{E}\left[\sum_{i\in P\cap S} \eta_i^2\right] \leq \Xi \tag{90}$$

Proof. We have

$$\sum_{i\in P\cap S} \eta_i^2 \leq \sum_{i\in P\cap S} \left(\eta_i^2 - \mathbb{V}\left(\eta_i\right)\right) + \mathbb{V}\left(\eta_i\right) \tag{91}$$

■

A.2 Polynomial Kernel

The polynomial kernel is given by

$$K(\mathbf{x}_1, \mathbf{x}_2) = \left(\mathbf{x}_1^\top \mathbf{x}_2 + C\right)^p \tag{92}$$

The feature map for the polynomial kernel is given as

$$\phi_{\mathsf{poly}}(\mathbf{x}) = \left[x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, \ldots, x_1^p, \ldots, x_d^p, x_1x_2, \ldots, x_{d-1}x_d\right] \in \mathbb{R}^{d^p} \tag{93}$$

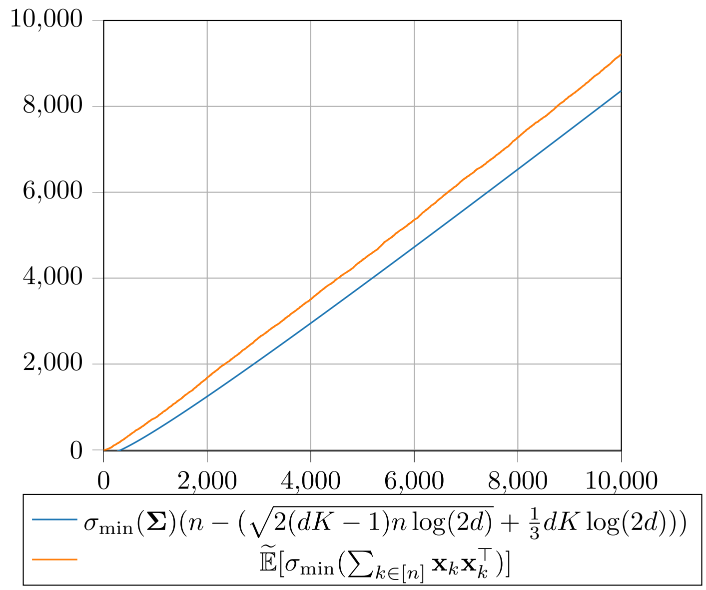

Figure 4: The bound for Lemma 30 compared to an average over 10 trials. We set $\mathbf{\Sigma} = \mathbf{I}$, $K = 4$, $d = 20$.

Lemma 32. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and let $\eta_1, \ldots, \eta_n \sim \mathcal{N}(0, \sigma^2)$ be sub-Gaussian. Given $C \in \mathbb{R}_+$, it then follows*

$$\mathbb{E}\left|\sum_{i=1}^{n} \eta_i \left(\|\mathbf{x}\|^2 + C\right)^{p/2}\right| \leq \Xi \tag{94}$$

Lemma 33. *Let $\mathbf{x}_1, \ldots, \mathbf{x}_n \sim \mathcal{N}_d(\mathbf{0}, \mathbf{I})$ such that $n > d^p$. It then follows*

$$\sigma_{\min}\left(\sum_{i=1}^{n} \phi_{\mathsf{poly}}(\mathbf{x}_i)\phi_{\mathsf{poly}}(\mathbf{x}_i)^\top\right) \geq \Xi \tag{95}$$

A.3 Gaussian Kernel

The gaussian kernel is given for $\gamma > 0$ by

$$K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\gamma \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2\right) \tag{96}$$

Lemma 34. *Let $\eta_1, \ldots, \eta_n \sim \mathcal{N}(0, \sigma^2)$ be sub-Gaussian. Then it follows*

$$\mathbb{E}\left|\sum_{i=1}^{n} \eta_i\right| \leq \Xi \tag{97}$$

We will utilize the Random Fourier Features (RFF) representation [27]. Let $\mathbf{w}_1, \ldots, \mathbf{w}_d \sim \mathcal{N}_d(\mathbf{0}, \frac{2}{\gamma}\mathbf{I})$, then the RFF reprsentation is given by

$$\phi_{\mathsf{RFF}}(\mathbf{x}) = \frac{1}{\sqrt{d}}\left[\cos\left(\mathbf{w}_1^\top\mathbf{x}\right), \sin\left(\mathbf{w}_1^\top\mathbf{x}\right), \cdots, \cos\left(\mathbf{w}_d^\top\mathbf{x}\right), \sin\left(\mathbf{w}_d^\top\mathbf{x}\right)\right] \tag{98}$$

The key idea behind the Randomized Fourier Features is

$$\mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_1)^\top\phi_{\mathsf{RFF}}(\mathbf{x}_2)\right] = \exp\left(-\gamma \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2\right) \tag{99}$$

We will first derive a relation between $\mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_1)^\top\phi_{\mathsf{RFF}}(\mathbf{x}_2)\right]$ and $\mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_1)\right]^\top \mathbb{E}\left[\phi_{\mathsf{RFF}}(\mathbf{x}_2)\right]$.

D Necessary Lemmas

Lemma 33 (MGF of Sub-Exponential Random Variable)**.** *Let x be a Sub-exponetial variable, then we have*

$$\mathbb{E}\left[\exp\left(\lambda x\right)\right] \leq \exp\left(C\lambda^2\right) \tag{135}$$

Theorem 34 (Matrix Chernoff, [**?**])**.** *Let $\mathbf{X}_k$ be a sequence of independent, random, self-adjoint matrices with dimension d s.t.*

$$\mathbf{X}_k \succcurlyeq \mathbf{0} \quad \textit{and} \quad \lambda_{\max}\left(\mathbf{X}_k\right) \leq R \quad \textit{almost surely} \tag{136}$$

Define

$$\mu_{\min} \triangleq \lambda_{\min}\left(\sum_k \mathbb{E}\mathbf{X}_k\right) \tag{137}$$

Then for $\delta \in [0,1]$

$$\mathbb{P}\left\{\lambda_{\min}\left(\sum_k \mathbf{X}_k\right) \geq (1-\delta)\mu_{\min}\right\} \leq d \cdot \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu_{\min}/R} \tag{138}$$

Theorem 35 (Matrix Bernstein, [33])**.** *Let $\mathbf{X}_k$ be a sequence of independent, random, self-adjoing matrices with dimension d s.t.*

$$\mathbb{E}\mathbf{X}_k = \mathbf{0} \tag{139}$$