CS 229
Project Final Report
Fast optimization of functions of inner products via Newton-Stein Method

Murat A. Erdogdu

erdogdu@stanford.edu
Department of Statistics, Stanford University

December 11, 2015

1 Introduction

In this paper, we consider the problem of minimizing a sum of n functions

$$\underset{\theta}{\text{minimize}}\ f(\theta) := \frac{1}{n}\sum_{i=1}^{n} f_i(\theta) \tag{1}$$

through projected iterations onto the compact and bounded parameter set $\mathcal{C} \subset \mathbb{R}^p$, where $f_i : \mathbb{R} \to \mathbb{R}$ are 4-times almost everywhere differentiable convex functions, $\theta \in \mathcal{C}$ is the parameter. We assume that the functions f_i's can be represented as inner products of the parameter θ and feature vectors $x_i \in \mathbb{R}^p$. That is,

$$f_i(\theta) = \psi(\langle x_i, \theta \rangle), \tag{2}$$

for some function θ.

This problem is commonly encountered in Machine Learning literature where ψ is chosen to be a suitable loss function quantifying the misfit, such as ℓ_p-loss or Hinge-p loss. In this paper, we focus on the regime where the number of observations/samples n, is much larger than the dimension of the data p, i.e., $n \gg p \gg 1$. In the target regime, stochastic algorithms are quite popular as their per-iteration cost is $\mathcal{O}(p)$ which is independent of the number of observations n. Such methods use a gradient (or sub-gradient) of a single, randomly selected observation to update the current iterate [Bottou, 2010]. Though they have negligible per-iteration cost, the convergence rate of these methods might be extremely slow. There are several extensions of the classical stochastic descent algorithms, providing significant improvement and/or stability [Bottou, 2010, Duchi et al., 2011, Schmidt et al., 2013].

Batch algorithms, on the other hand, enjoy faster convergence rates, though their per-iteration cost may be prohibitive. In particular, second order methods attain quadratic convergence rate, but constructing the exact Hessian matrix generally requires excessive amount of computation. Many algorithms aim at forming an approximate, cost-efficient scaling matrix. In particular, this idea lies at the core of Quasi-Newton methods [Bishop, 1995].

An alternative approach to construct an approximate Hessian matrix makes use of sub-sampling techniques [Martens, 2010, Byrd et al., 2011, Vinyals and Povey, 2012, Erdogdu and Montanari, 2015]. Many contemporary learning methods rely on sub-sampling as it is simple and it provides significant boost over the first order methods. Further improvements through conjugate gradient methods and Krylov sub-spaces are available.

In this paper, we focus on the iterations of the form (projection is omitted here for simplicity),

$$\theta^{t+1} \leftarrow \theta^t - \gamma \mathbf{Q}^t \nabla_\theta f(\theta^t)\,, \tag{3}$$

where $\mathbf{Q}^t$ is a suitable scaling matrix updated at each iteration, $\nabla_\theta f(\theta)$ is the gradient of the function f and γ is the step size. Our focus is to construct a suitable and cost-efficient sequence of scaling matrices $\{\mathbf{Q}^t\}_{t>0}$ that provides sufficient curvature information as in Quasi-Newton methods.

In a recent paper [Erdogdu, 2015], authors proposed a method called "Newton-Stein Method" for training Generalized Linear Models (GLMs). The method is based on a Stein-type lemma, which provides a fast update per each iteration. The simple idea is to remove the dependence of the matrices involved in the computation of $\mathbf{Q}^t$ to the iterates. This ways, such matrices can be computed only once and used throughout the iterations without requiring update. Even though the experiments are convincing, the novelty is limited to GLMs such as linear/logistic regression.

In this paper, we generalize Newton-Stein method to functions of inner products which broadens the range of applicability of the suggested method. The method will again rely on the same argument based on Stein's lemma, and following [Erdogdu and Montanari, 2015], we will introduce further improvements through sub-sampling and eigenvalue thresholding. This way, the algorithm will be applicable to a wide variety of problems such as Support vector machines with Hinge-p loss, linear programing with log-barriers or even non-convex problems such as Neural Networks etc.

2 Proposed Algorithm: Generalized Newton-Stein Method

The task of constructing an approximate Hessian can be viewed as an estimation problem. Assume for simplicity that f_i's denote a loss function between the linear prediction and the observation, i.e. $f_i(\theta) = \psi(y_i, \langle x_i, \theta\rangle)$. For simplicity, we will mostly drop the dependence on the observations and simply write $f_i(\theta) = \psi(\langle x_i, \theta\rangle)$.

Assuming that the features x_i's are i.i.d. random vectors with bounded support, the Hessian of the aforementioned minimization problem has the following form

$$\left[\mathbf{Q}^t\right]^{-1} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T \psi^{(2)}(\langle x_i, \theta\rangle) \approx \mathbb{E}[xx^T\psi^{(2)}(\langle x, \theta\rangle)]\,, \tag{4}$$

where $\psi^{(2)}$ denotes the second derivative of the function f. We observe that $\left[\mathbf{Q}^t\right]^{-1}$ is just a sum of i.i.d. matrices. Hence, the true Hessian is nothing but a sample mean estimator to its expectation. Another natural estimator would be the sub-sampled Hessian method suggested by [Martens, 2010, Byrd et al., 2011, Erdogdu and Montanari, 2015]. Similarly, our goal is to propose an appropriate estimator that is also computationally efficient.

Following the idea in [Erdogdu, 2015], we use the following Stein-type lemma to derive an efficient estimator to the expectation of Hessian.

Lemma 2.1 (Stein-type lemma-[Erdogdu, 2015])**.** *Assume that $x \sim \mathsf{N}_p(0, \mathbf{\Sigma})$ and $\theta \in \mathbb{R}^p$ is a constant vector. Then for any function $\psi : \mathbb{R} \to \mathbb{R}$ that is twice* "weakly" *differentiable, we have*

$$\mathbb{E}[xx^T\psi(\langle x, \theta\rangle)] = \mathbb{E}[\psi(\langle x, \theta\rangle)]\mathbf{\Sigma} + \mathbb{E}[\psi^{(2)}(\langle x, \theta\rangle)]\mathbf{\Sigma}\theta\theta^T\mathbf{\Sigma}\,. \tag{5}$$

The proof of Lemma 2.1 is straightforward and can be found in [Erdogdu, 2015]. In Eqs. 4 and 5, we realize that the true Hessian is also an estimator to the quantity on provided by Lemma 2.1. Therefore, we can instead come up with an estimator for the right hand side of Eq. 5. We notice that the right hand side of Eq.(5) becomes a rank-1 update to the first term. Hence, its inverse can be computed with $\mathcal{O}(p^2)$ cost. Quantities that change at each iteration are the ones that depend on θ, i.e.,

$$\mu_2(\theta) = \mathbb{E}[\psi^{(2)}(\langle x, \theta\rangle)], \quad \text{and} \quad \mu_4(\theta) = \mathbb{E}[\psi^{(4)}(\langle x, \theta\rangle)].$$

$\mu_2(\theta)$ and $\mu_4(\theta)$ are scalar quantities and can be estimated by their corresponding sample means $\hat{\mu}_2(\theta) = \frac{1}{n}\sum_{i=1}^n \psi^{(2)}(\langle x_i, \theta\rangle)$ and $\hat{\mu}_4(\theta) = \frac{1}{n}\sum_{i=1}^n \psi^{(4)}(\langle x_i, \theta\rangle)$, with only $\mathcal{O}(np)$ computation. We emphasize that sub-sampling can be also used at this stage. But since we consider batch updates, cost of computing the gradient is $\mathcal{O}(np)$. Therefore, using sub-sampling at this stage of the algorithm will not provide huge impact in terms of order of magnitude.

To complete the estimation task suggested by Eq. (5), we need an estimator for the covariance matrix $\mathbf{\Sigma}$. A natural estimator is the sample mean where, we only use a sub-sample $S \subset [n]$ so that the cost is reduced to $\mathcal{O}(|S|p^2)$ from $\mathcal{O}(np^2)$. Sub-sampling based sample mean estimator is denoted by $\widehat{\mathbf{\Sigma}}_S = \sum_{i\in S} x_i x_i^T/|S|$, which is widely used in large-scale problems [Vershynin, 2010]. We highlight the fact that Lemma 2.1 replaces Newton Method's (NM) $O(np^2)$ per-iteration cost with a one-time cost of $O(np^2)$. We further use sub-sampling to reduce this one-time cost to $O(|S|p^2)$.

In general, important curvature information is contained in the largest few spectral features. Following [Erdogdu and Montanari, 2015], we take the largest r eigenvalues of the sub-sampled covariance estimator, setting rest of them to $(r+1)$-th eigenvalue. This operation helps denoising and would require $\mathcal{O}(rp^2)$ computation. This eigenvalue thresholding operation is denoted by ζ_r below.

Notice that the updates in Eq. (5) are based on rank-1 matrix additions. Hence, we can simply use a matrix inversion formula to derive an explicit equation. This formulation would impose another inverse operation on the covariance estimator. Since the covariance estimator is also based on rank-r approximation, one can utilize the low-rank inversion formula again. We emphasize that this operation is performed once. Therefore, instead of NM's per-iteration cost of $\mathcal{O}(p^3)$ due to inversion, the proposed algorithm requires $\mathcal{O}(p^2)$ per-iteration and a one-time cost of $\mathcal{O}(rp^2)$. Assuming that our algorithm and NM converge in T_1 and T_2 iterations respectively, the overall complexity of our algorithm is $\mathcal{O}\left(npT_1 + p^2T_1 + (|S| + r)p^2\right) \approx \mathcal{O}\left(npT_1 + p^2T_1 + |S|p^2\right)$ whereas that of NM is $\mathcal{O}\left(np^2T_2 + p^3T_2\right)$.

Even though Proposition 2.1 assumes that the covariates are multivariate Gaussian random vectors, the only assumption we will make on the covariates is that they have bounded support, which covers a wide class of random variables. We note that bounded support assumption can be further relaxed to a more general case called *sub-gaussian* random vectors, but this will not be pursued here. The main reason that Newton-Stein method works for general distributions is a consequence of the fact that the proposed estimator in Eq. (5) relies on the distribution of x only through inner products of the form $\langle x, v\rangle$, which in turn results in approximate normal distribution due to the central limit theorem when p is sufficiently large. We will discuss this phenomenon in detail in the final report. Combining the aforementioned arguments, we may write the following for an Hessian estimator:

$$[\mathbf{Q}^t]^{-1} = \hat{\mu}_2(\theta^t)\zeta_r(\widehat{\mathbf{\Sigma}}_S) + \hat{\mu}_4(\theta^t)\zeta_r(\widehat{\mathbf{\Sigma}}_S)\theta^t[\theta^t]^T\zeta_r(\widehat{\mathbf{\Sigma}}_S).$$

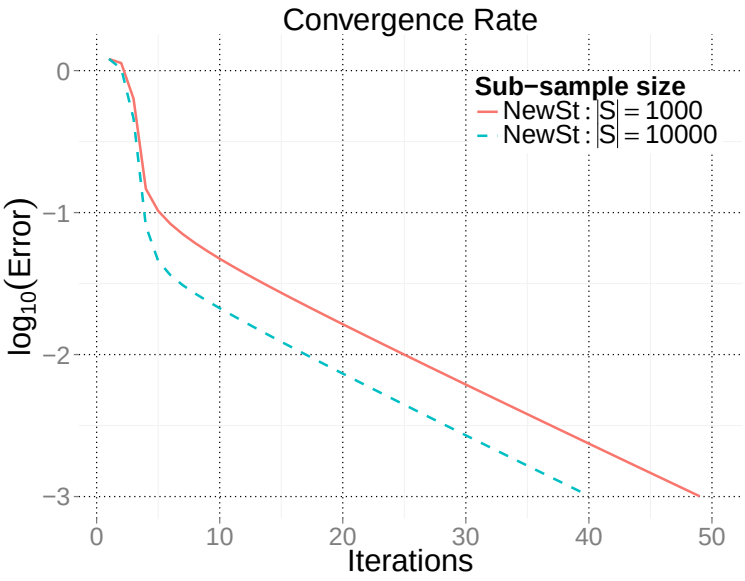

Figure 1: Composite convergence over two different sub-sampling sizes.

As mentioned before, the second term on the right hand side is just a rank-1 update to the first term. Therefore, we can invoke a fast inversion formula and obtain the the scaling matrix

$$\mathbf{Q}^t = \frac{1}{\hat{\mu}_2(\theta^t)} \left[\zeta_r(\widehat{\mathbf{\Sigma}}_S)^{-1} - \frac{\theta^t[\theta^t]^T}{\hat{\mu}_2(\theta^t)/\hat{\mu}_4(\theta^t) + \langle \zeta_r(\widehat{\mathbf{\Sigma}}_S)\theta^t, \theta^t \rangle} \right]. \tag{6}$$

3 Current theoretical results and future work

The main contribution of this work is the generalization of the bound given only for GLMs by [Erdogdu, 2015] to functions of inner products. The following theorem is our main result:

Theorem 3.1. *Assume that the function ψ has bounded and Lipschitz continuous second and fourth derivatives. Let the features $x_1, x_2, ..., x_n$ be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$ with*

$$\mathbb{E}[x_i] = 0 \qquad \textit{and} \qquad \mathbb{E}\big[x_i x_i^T\big] = \mathbf{\Sigma},$$

where $\mathbf{\Sigma}$ follows the r-spiked model.

If $n, |S|$ and p are sufficiently large, then there exist constants c, c_1, c_2, c_3 depending on the properties of the covariates and ψ such that with probability at least $1 - c/p^2$, we have

$$\big\|\hat{\theta}^{t+1} - \theta_*\big\|_2 \leq \tau_1 \big\|\hat{\theta}^t - \theta_*\big\|_2 + \tau_2 \big\|\hat{\theta}^t - \theta_*\big\|_2^2, \tag{7}$$

where the coefficients τ_1 and τ_2 are deterministic constants defined as

$$\tau_1 = c_1 \mathfrak{D}(x, z) + c_2 \sqrt{\frac{p}{\min\{|S|, n/\log(n)\}}}, \qquad \tau_2 = c_3,$$

and $\mathfrak{D}(x, z)$ is a probability metric defined in Lemma (B.2)

The above theorem provides a per-step bound which can be further used to obtain bounds on number of iterations. However, the main importance of the above result is that it explains the dependence of convergence behavior to the data dimensions quite well. In particular, we first notice that as the squared term would dominate, the convergence will start as quadratic at start. Therefore, the initial convergence will be determined by the convergence coefficient τ_2. As the iterations proceed to the correct minimizer, the squared term will become small and therefore, the linear term will dominate. Therefore, the slope of the linear term will be determined by τ_1. This behavior can be observed in Figure 1 where we notice that sub-sampling size also has a major effect in the convergence rate. That is, larger sub-sample size gives a longer quadratic rate of convergence. However, this imposes a trade-off between the faster convergence rate and the per-iteration cost.

We notice that the coefficient of the quadratic term is just a constant. However, the coefficient of the linear term is composed of three components: The first term in τ_1 quantifies how accurate the Gaussian approximation was. We expect this term to be small as the dependence of the parameter and the feature vectors are linear, which will invoke a central limit theorem given that the θ is well-spread.

4 Future work towards a complete paper

In this work, we generalized the Newton-Stein Method to more general minimization problems which was previously proposed for GLMs. We state the following probable future work to get a complete paper.

- An upper bound on the number of iterations can be easily obtained using the properties of the composite convergence. The upper bound will have two terms: 1- one from quadratically converging term, 2- one from the linearly converging term.
- The distributional assumptions on the covariates, i.e. bounded support, can be relaxed to more general class of distributions such as sub-gaussian. Even though this requires significant amount of manipulations, the proof idea is quite similar.
- We can demonstrate the value and applicability of the generalized Newton-Stein method on examples such as *Linear Programing with logarithmic bariers*, *Linear Support Vector Machines with Hinge-p loss* etc. We also note that when applied to a GLM problem, generalized Newton-Stein method reduces to the classical Newton-Stein.
- Extensive numerical studies are needed to compare the proposed algorithm to well-known optimization methods. We expect to see significant improvement over Quasi-Newton methods such as BFGS, L-BFGS as they only rely on gradients and iterates to construct an approximate scaling matrix.

References

[Bishop, 1995] Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, Inc., NY, USA.

[Bottou, 2010] Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *COMPSTAT*, pages 177–186.

[Byrd et al., 2011] Byrd, R. H., Chin, G. M., Neveitt, W., and Nocedal, J. (2011). On the use of stochastic hessian information in optimization methods for machine learning. *SIAM Journal on Optimization*, 21(3):977–995.

[Duchi et al., 2011] Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159.

[Erdogdu, 2015] Erdogdu, M. A. (2015). Newton-Stein Method: A second order method for GLMs via Stein's lemma. In *NIPS*.

[Erdogdu and Montanari, 2015] Erdogdu, M. A. and Montanari, A. (2015). Convergence rates of sub-sampled Newton methods. In *NIPS*.

[Martens, 2010] Martens, J. (2010). Deep learning via hessian-free optimization. In *ICML*, pages 735–742.

[Schmidt et al., 2013] Schmidt, M., Roux, N. L., and Bach, F. (2013). Minimizing finite sums with the stochastic average gradient. *arXiv preprint arXiv:1309.2388*.

[Vershynin, 2010] Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*.

[Vinyals and Povey, 2012] Vinyals, O. and Povey, D. (2012). Krylov Subspace Descent for Deep Learning. In *AISTATS*.

A Proof of the theorem

We will provide the proofs of the Theorem 3.1. Matrix concentration results in this section are mostly based on the covering net argument provided in [Vershynin, 2010]. Similar results for matrix forms can also be obtained through different techniques such as *chaining* as well. On the set $\mathcal{E}$, we write,

$$\begin{aligned}\hat{\theta}^t - \theta_* - \gamma \mathbf{Q}^t \nabla_\theta f(\hat{\theta}^t) &= \hat{\theta}^t - \theta_* - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi (\hat{\theta}^t - \theta_*), \\ &= \left(I - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi \right) (\hat{\theta}^t - \theta_*) \,.\end{aligned}$$

Since the projection $\mathcal{P}_\mathcal{C}$ can only decrease the ℓ_2 distance, we obtain

$$\|\hat{\theta}^{t+1} - \theta_*\|_2 \leq \left\| I - \gamma \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi \right\|_2 \|\hat{\theta}^t - \theta_*\|_2. \tag{8}$$

The governing term (with $\gamma = 1$) that determines the convergence rate (the first term on the right hand side) can be bounded as

$$\left\| I - \mathbf{Q}^t \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi \right\|_2 \leq \left\| [\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi \right\|_2 \|\mathbf{Q}^t\|_2.$$

We define the following,

$$\mathfrak{E}(\theta) = \mathbb{E}\left[\psi^{(2)}(\langle x, \theta \rangle)\right] \boldsymbol{\Sigma} + \mathbb{E}\left[\psi^{(4)}(\langle x, \theta \rangle)\right] \boldsymbol{\Sigma} \theta \theta^T \boldsymbol{\Sigma}$$

Note that for a function f, $\mathbb{E}[f(\langle x, \theta \rangle)] = h(\theta)$ is a function of θ. With a slight abuse of notation, we write $\mathbb{E}[f(\langle x, \hat{\theta} \rangle)] = h(\hat{\theta})$ as a random variable. We have

$$\begin{aligned}\left\| [\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi \right\|_2 &\leq \left\| [\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\theta}^t) \right\|_2 \\ &+ \left\| [\mathbb{E}[x x^T \psi^{(2)}(\langle x, \hat{\theta}^t \rangle)] - \mathfrak{E}(\hat{\theta}^t) \right\|_2 \\ &+ \left\| \int_0^1 \boldsymbol{\nabla}^2_\theta \ell(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi - \mathbb{E}\left[x x^T \int_0^1 \psi^{(2)}(\langle x, \theta_* + \xi(\hat{\theta}^t - \theta_*) \rangle) d\xi \right] \right\|_2 \\ &+ \left\| \mathbb{E}[x x^T \psi^{(2)}(\langle x, \hat{\theta}^t \rangle)] - \mathbb{E}\left[x x^T \int_0^1 \psi^{(2)}(\langle x, \theta_* + \xi(\hat{\theta}^t - \theta_*) \rangle) d\xi \right] \right\|_2 .\end{aligned}$$

We find an upper bound for the terms on the right hand side. In order to accomplish this, we state a lemma for each term in Section B.

Using Lemmas B.1, B.2, B.3 and B.4, the right hand side can be bounded above by

$$\begin{aligned}&\left\| [\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*)) d\xi \right\|_2 \\ &\leq c_1 \sqrt{\frac{p}{\min\{|S|, n/\log(n)\}}} + \mathfrak{D}(x, z) + c_2 \sqrt{\frac{p}{n} \log(n)} + c_3 \|\hat{\theta}^t - \theta_*\|_2\end{aligned}$$

Lastly, using Lemma B.5, we obtain

$$\|\mathbf{Q}^t\|_2 \leq \kappa.$$

Combining the above results we get

$$\left\|[\mathbf{Q}^t]^{-1} - \int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*))d\xi\right\|_2 \tag{9}$$

$$\kappa\left\{c_1\sqrt{\frac{p}{\min\{|S|, n/\log(n)\}}} + \mathfrak{D}(x,z) + c_2\sqrt{\frac{p}{n}\log(n)} + c_3\|\hat{\theta}^t - \theta_*\|_2\right\}. \tag{10}$$

Finally, we combine the above inequalities and obtain our main result:

$$\begin{aligned}\|\hat{\theta}^{t+1} - \theta_*\|_2 \leq& \left\|I - \gamma\mathbf{Q}^t\int_0^1 \boldsymbol{\nabla}^2_\theta f(\theta_* + \xi(\hat{\theta}^t - \theta_*))d\xi\right\|_2 \|\hat{\theta}^t - \theta_*\|_2 \\ \leq& \kappa\left\{c_1\sqrt{\frac{p}{\min\{|S|, n/\log(n)\}}} + \mathfrak{D}(x,z) + c_2\sqrt{\frac{p}{n}\log(n)} + c_3\|\hat{\theta}^t - \theta_*\|_2\right\}\|\hat{\theta}^t - \theta_*\|_2.\text{`}\end{aligned}$$

B Main Lemmas

Lemma B.1 ([Erdogdu, 2015])**.** *There exist constants c, C such that, with probability at least $1 - c/p^2$,*

$$\left\|[\mathbf{Q}^t]^{-1} - \mathfrak{E}(\hat{\theta}^t)\right\|_2 \leq C\sqrt{\frac{p}{\min\{|S|, n/\log(n)\}}},$$

where the constants depend on K, B and the radius R.

Based on the second and fourth derivatives of cumulant generating function, we define the following function classes:

$$\mathcal{H}_1 = \left\{h(x) = \psi^{(2)}(\langle x, \theta\rangle) : \theta \in \mathcal{C}\right\}, \qquad \mathcal{H}_2 = \left\{h(x) = \psi^{(4)}(\langle x, \theta\rangle) : \theta \in \mathcal{C}\right\},$$

$$\mathcal{H}_3 = \left\{h(x) = \langle v, x\rangle^2\psi^{(2)}(\langle x, \theta\rangle) : \theta \in \mathcal{C}, \|v\|_2 = 1\right\},$$

Lemma B.2. *The bias term is can be upper bounded by the probability metric $\mathfrak{D}(x,z)$ for a gaussian random vector z, i.e.,*

$$\left\|\mathbb{E}[xx^T\psi^{(2)}(\langle x, \hat{\theta}^t\rangle)] - \mathfrak{E}(\hat{\theta}^t)\right\|_2 \leq d_{\mathcal{H}_3}(x,z) + \|\boldsymbol{\Sigma}\|_2\ d_{\mathcal{H}_1}(x,z) + \|\boldsymbol{\Sigma}\|_2^2 R^2\ d_{\mathcal{H}_2}(x,z).$$

Lemma B.3 ([Erdogdu, 2015])**.** *There exist constants c_1, c_2, c_3 depending on K, B, L and R such that, with probability at least $1 - c_2e^{-c_3 p}$*

$$\left\|\frac{1}{n}\sum_{i=1}^n x_i x_i^T\int_0^1 \psi^{(2)}(\langle x_i, \theta_* + \xi(\hat{\theta}^t - \theta_*)\rangle)d\xi - \mathbb{E}\left[xx^T\int_0^1 \psi^{(2)}(\langle x, \theta_* + \xi(\hat{\theta}^t - \theta_*)\rangle)d\xi\right]\right\|_2 \leq c_1\sqrt{\frac{p}{n}\log(n)}.$$

Lemma B.4. *There exists a constant C depending on K and L such that,*

$$\left\|\mathbb{E}[xx^T\psi^{(2)}(\langle x,\hat{\theta}^t\rangle)]-\mathbb{E}\left[xx^T\int_0^1\psi^{(2)}(\langle x,\theta_*+\xi(\hat{\theta}^t-\theta_*)\rangle)d\xi\right]\right\|_2\leq C\|\hat{\theta}^t-\theta_*\|_2.$$

Proof. First, we apply the Fubini's theorem, and obtain

$$\begin{aligned}&\left\|\mathbb{E}[xx^T\psi^{(2)}(\langle x,\hat{\theta}^t\rangle)]-\mathbb{E}\left[xx^T\int_0^1\psi^{(2)}(\langle x,\theta_*+\xi(\hat{\theta}^t-\theta_*)\rangle)d\xi\right]\right\|_2,\\&=\left\|\int_0^1\mathbb{E}\left[xx^T\left\{\psi^{(2)}(\langle x,\hat{\theta}^t\rangle)-\psi^{(2)}(\langle x,\theta_*+\xi(\hat{\theta}^t-\theta_*)\rangle)\right\}\right]d\xi\right\|_2,\end{aligned}$$

We take the integration out, and obtain

$$\begin{aligned}&\int_0^1\left\|\mathbb{E}\left[xx^T\left\{\psi^{(2)}(\langle x,\hat{\theta}^t\rangle)-\psi^{(2)}(\langle x,\theta_*+\xi(\hat{\theta}^t-\theta_*)\rangle)\right\}\right]\right\|_2d\xi,\\&\leq\int_0^1\left\|\mathbb{E}\left[xx^TL|\langle x,(1-\xi)(\hat{\theta}^t-\theta_*)\rangle|\right]\right\|_2d\xi,\\&\leq\mathbb{E}\left[\|x\|_2^3\|\hat{\theta}^t-\theta_*\|_2\right]L\int_0^1(1-\xi)d\xi,\\&=\frac{LK^{3/2}}{2}\|\hat{\theta}^t-\theta_*\|_2.\end{aligned}$$

□

Lemma B.5 ([Erdogdu, 2015])**.** *If n, $|S|$ are sufficiently large relative to p, we have*

$$\left\|\mathbf{Q}^t\right\|_2=\left\|\frac{1}{\hat{\mu}_2(\hat{\theta}^t)}\left[\zeta_r(\widehat{\mathbf{\Sigma}}_S)^{-1}-\frac{\hat{\theta}^t[\hat{\theta}^t]^T}{\hat{\mu}_2(\hat{\theta}^t)/\hat{\mu}_4(\hat{\theta}^t)+\langle\zeta_r(\widehat{\mathbf{\Sigma}}_S)\hat{\theta}^t,\hat{\theta}^t\rangle}\right]\right\|_2\leq\kappa.$$

for some constant κ depending on the properties of the functions f_i.

C Hoeffding-type inequalities

Lemma C.1. *Let $x_i\in\mathbb{R}^p$, for $i=1,2,...,n$, be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$, with mean 0, and covariance matrix $\mathbf{\Sigma}$. Also let $f:\mathbb{R}\to\mathbb{R}$ be a uniformly bounded function such that for some $B>0$, we have $\|f\|_\infty<B$ and f is Lipschitz continuous with constant L. Then, there exist constants c_1,c_2,c_3 such that*

$$\mathbb{P}\left(\sup_{\theta\in B_n(R)}\left|\frac{1}{n}\sum_{i=1}^n f(\langle x_i,\theta\rangle)-\mathbb{E}[f(\langle x,\theta\rangle)]\right|>c_1\sqrt{\frac{p\log(n)}{n}}\right)\leq c_2e^{-c_3p},$$

where the constants depend only on the bound B and radii R and $\sqrt{K}$.

We skip the proof of above lemma as it is similar to the that of below lemma.

Lemma C.2. *Let $x_i \in \mathbb{R}^p$, for $i = 1, ..., n$, be i.i.d. random vectors supported on a ball of radius $\sqrt{K}$, with mean 0, covariance matrix $\mathbf{\Sigma}$. Also let $f : \mathbb{R} \to \mathbb{R}$ be a uniformly bounded function such that for some $B > 0$, we have $\|f\|_\infty < B$ and f is Lipschitz continuous with constant L. Then, for $v \in S^{p-1}$, there exist constants c_1, c_2, c_3 such that*

$$\mathbb{P}\left(\sup_{\theta \in B_p(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| > c_1\sqrt{\frac{p\log(n)}{n}}\right) \leq c_2 e^{-c_3 p},$$

where the constants depend only on the bound B and radii R and $\sqrt{K}$.

Proof of Lemma C.2. As in the proof of Lemma C.1, we start by using the Lipschitz property of the function f, i.e., $\forall \theta, \theta' \in B_p(R)$,

$$\begin{aligned}\|f(\langle x, \theta\rangle)\langle x, v\rangle^2 - f(\langle x, \theta'\rangle)\langle x, v\rangle^2\|_2 \leq & L\|x\|_2^3\|\theta - \theta'\|_2, \\ \leq & LK^{1.5}\|\theta - \theta'\|_2.\end{aligned}$$

For a net T_Δ, $\forall \theta \in B_p(R)$, $\exists \theta' \in T_\Delta$ such that right hand side of the above inequality is smaller than $\Delta L\sqrt{K}$. Then, we can write

$$\begin{aligned}\left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| \leq & \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta'\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta'\rangle)\langle x, v\rangle^2] \right| \\ & + 2\Delta LK^{1.5}. \end{aligned} \tag{11}$$

This time, we choose

$$\Delta = \frac{\epsilon}{4LK^{1.5}},$$

and take the supremum over the corresponding feasible θ-sets on both sides,

$$\begin{aligned}&\sup_{\theta \in B_n(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| \\ &\leq \max_{\theta \in \mathcal{T}_\Delta} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| + \frac{\epsilon}{2}.\end{aligned}$$

Now, since we have $\|f\|_\infty \leq B$ and for fixed θ and v, $i = 1, 2, ..., n$, $f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2$ are i.i.d. random variables. By the Hoeffding's concentration inequality, we have

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| > \epsilon/2\right) \leq 2\exp\left(-\frac{n\epsilon^2}{2B^2K^2}\right).$$

Using Eq. (11) and the above result combined with the union bound, we easily obtain

$$\begin{aligned}&\mathbb{P}\left(\sup_{\theta \in B_n(R)} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| > \epsilon\right) \\ &\leq \mathbb{P}\left(\max_{\theta \in \mathcal{T}_\Delta} \left| \frac{1}{n}\sum_{i=1}^{n} f(\langle x_i, \theta\rangle)\langle x_i, v\rangle^2 - \mathbb{E}[f(\langle x, \theta\rangle)\langle x, v\rangle^2] \right| > \epsilon/2\right) \\ &\leq 2|\mathcal{T}_\Delta|\exp\left(-\frac{n\epsilon^2}{2B^2K^2}\right),\end{aligned}$$

where $\Delta = \epsilon/4LK^{1.5}$. By standard ϵ-net arguments over a ball of radius R in p dimensions we have

$$|\mathcal{T}_\Delta| \leq \left(\frac{R\sqrt{p}}{\Delta}\right)^p = \left(\frac{R\sqrt{p}}{\epsilon/4LK^{1.5}}\right)^p.$$

As before, we require that the right hand side of above inequality gets a decay with rate $\mathcal{O}(p)$. By straightforward algebraic manipulations, we obtain that ϵ should be

$$\epsilon = \sqrt{\frac{B^2K^2p}{n}\log\left(\frac{16L^2R^2K^3n}{B^2}\right)} = \mathcal{O}\left(\sqrt{\frac{p\log(n)}{n}}\right),$$

which completes the proof. □