Hebbian learning inspired estimation of the linear regression parameters from queries

Johannes Schmidt-Hieber* and Wouter M. Koolen†

September 26, 2023

Abstract

Local learning rules in biological neural networks (BNNs) are commonly referred to as Hebbian learning. [26] links a biologically motivated Hebbian learning rule to a specific zeroth-order optimization method. In this work, we study a variation of this Hebbian learning rule to recover the regression vector in the linear regression model. Zeroth-order optimization methods are known to converge with suboptimal rate for large parameter dimension compared to first-order methods like gradient descent, and are therefore thought to be in general inferior. By establishing upper and lower bounds, we show, however, that such methods achieve near-optimal rates if only queries of the linear regression loss are available. Moreover, we prove that this Hebbian learning rule can achieve considerably faster rates than any non-adaptive method that selects the queries independently of the data.

Keywords: Biological neural networks, derivative-free methods, linear regression model, minimax estimation, sequential estimation, zeroth-order optimization.

MSC 2020: Primary: 62L20; secondary: 62J05

1 Introduction

While considerable efforts have been dedicated to develop the theoretical foundations underlying artificial neural networks (ANNs), a theory for learning in biological neural networks (BNNs) remains largely unexplored.

ANNs were designed to mimic BNNs but there are distinct differences in terms of the network structure and the learning. ANNs are deterministic and pass real numbers through the network whereas BNNs are stochastic and biological neurons send so-called spike trains. The information in BNNs is decoded in the spike times, the moments when the neurons fire/discharge. The brain updates the weights locally based on the state of the neurons that it connects. There

*University of Twente
†University of Twente and CWI

are various quantitative and qualitative updating rules and they are commonly referred to as Hebbian learning. Specifically, spike-time dependent plasticity rules are Hebbian learning rules based on local spike times. It has been widely acknowledged in previous work that Hebbian learning in BNNs cannot perform gradient descent [10, 20, 29]. The rationale behind is that updating one parameter in a gradient descent scheme requires to evaluate a partial derivative that depends on the other parameters. This cannot be implemented by a local learning rule.

The recent work [26] derives a specific derivative-free optimization method that captures key characteristics of the spike time dependent plasticity rule underlying Hebbian learning in BNNs. For mathematical tractability, we consider here a slight variation of this scheme.

We study this iterative scheme in a scenario where instead of the full data, one can only query the linear regression model, see Section 2.2 for a definition and discussion. Based on k rounds of querying, the aim of the method is to recover the d-dimensional regression vector. Estimation in this setting is non-trivial. In particular, the data are not informative enough to compute gradients and run gradient descent. A first contribution of this article is to derive a bound for the convergence rate of the biologically inspired gradient-free learning rule in the query model. It is moreover shown that up to a log-factor in the number of parameters d, the derived convergence rates matches the lower bound for sample sizes $k \gtrsim d^2 \log(d)$. The derived minimax lower bound is non-standard. The main obstacle is that sequential estimation procedures with queries depending on previous observations induces dependence among the data. As we do not constraint the queries, the induced dependence on the future data is hard to characterize and to control. Compared to the earlier lower bounds for adaptive sensing [2] and derivative-free stochastic optimization [27, 12], the main difficulty here is that the dependence on previous queries and the parameter appears not in the mean but in the variance of the data distribution.

Any method achieving the optimal rate in the linear model with queries needs to carefully exploit the information obtained from previous queries. Indeed, we show that naive methods that specify the queries independently of the data will converge with a slower rate in the dimension parameter d, see Corollary 1 for the precise statement. This should be contrasted with adaptive sensing where it is known that the adaptation to previously seen data can improve the rate by, at most, a log-factor [2].

The article is structured as follows. Section 2 provides more details on biological learning and the link with gradient-free optimization. After formally stating the query model, the convergence rates are given. A corresponding lower bound is stated in Section 3. Section 4 derives matching upper and lower bounds for the convergence rate in the case of non-adaptive queries. This enables us to quantify the gain in the convergence rate by integrating previously observed query values into the query strategy. Related literature is summarized in Section 5. Proofs are deferred to the Appendix.

Notation: For vectors the spectral norm coincides with the Euclidean norm, see Lemma 4. Therefore, we can denote both norms by $\|\cdot\|$. Matrix inequalities are taken with respective to the partial ordering of symmetric matrices (Loewner order).

2 Upper bounds

2.1 Hebbian learning inspired update rule

Working in the supervised learning framework, suppose we want to learn a d-dimensional parameter vector $\boldsymbol{\theta}$ from training data $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots$, with d-dimensional design vectors/inputs $\mathbf{X}_1, \mathbf{X}_2, \ldots$ and real-valued response variables $Y_1, Y_2, \ldots$ For our approach, it is sufficient to assume that the number of parameters and the number of covariates are the same, that is, $\boldsymbol{\theta}$ and $\mathbf{X}_1, \mathbf{X}_2, \ldots$ are all vectors of length d.

Whereas gradient descent based methods are ubiquitous in machine learning, BNNs rely on local updating rules and receive feedback through neurotransmitters such as dopamine. Anticipating to solve a task well reduces the amount of released neurotransmitter in the brain. In reward-based synaptic plasticity, the amount of released neurotransmitter is modelled as the difference between a realized loss based on the current task and the anticipated loss that predicts the current loss based on previously seen losses [14].

Zeroth-order optimization methods are derivative-free methods that use the evaluation of the loss but do not involve the gradient of the loss. Since weight updates in BNNs are based on evaluations of the loss, it seems natural to interpret the learning in BNNs as a specific zeroth-order method. [26] extracts from the spike-time dependent local updating of the weights in a BNN a global zeroth-order rule for learning the parameter vector $\boldsymbol{\theta}$. If $\boldsymbol{\theta}_k$ denotes the k-th update, $L(\boldsymbol{\theta}, \mathbf{X}_k, Y_k)$ is the loss for parameter $\boldsymbol{\theta}$ on the k-th training sample $(\mathbf{X}_k, Y_k)$, $\overline{L}_k$ is the anticipated loss in the k-th round based on previously seen losses, and $\mathbf{U}_k$ is a d-dimensional uniform random vector $\mathrm{Unif}([A-, A]^d)$, the biologically inspired update formula is

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \alpha_k \big(L(\boldsymbol{\theta}_{k-1} + \mathbf{U}_k, \mathbf{X}_k, Y_k) - \overline{L}_k\big)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big). \tag{1}$$

The expressions $e^{\pm \mathbf{U}_k}$ have to be understood componentwise, that is, if $\mathbf{U}_k = (U_{k1}, \ldots, U_{kd})^\intercal$, then $e^{\mathbf{U}_k} = (e^{U_{k1}}, \ldots, e^{U_{kd}})^\intercal$ and $e^{-\mathbf{U}_k} = (e^{-U_{k1}}, \ldots, e^{-U_{kd}})^\intercal$. The parameter A in the uniform distribution is a constant of the biological network but treated here as a hyperparameter of the optimization method. There is little guidance from the neuroscience on how to choose the anticipated loss $\overline{L}_k$. In [26] the $\overline{L}_k$ was taken as the loss from the previous round $L(\boldsymbol{\theta}_{k-2} + \mathbf{U}_{k-1}, \mathbf{X}_{k-1}, Y_{k-1})$. For mathematical tractability, we assign instead the value $L(\boldsymbol{\theta}_{k-1} + \mathbf{U}'_k, \mathbf{X}_k, Y_k)$ to $\overline{L}_k$, for an independently sampled and uniformly distributed random vector $\mathbf{U}'_k \sim \mathrm{Unif}([-A, A]^d)$. Thus, we will study in this work convergence of the zeroth-order method

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \alpha_k \big(L(\boldsymbol{\theta}_{k-1} + \mathbf{U}_k, \mathbf{X}_k, Y_k) - L(\boldsymbol{\theta}_{k-1} + \mathbf{U}'_k, \mathbf{X}_k, Y_k)\big)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big), \quad k = 1, \ldots \tag{2}$$

Theorem 1 in [26] shows that in expectation, this rule does approximately gradient descent. Working here with a slightly different anticipated loss does not change this result. Indeed, since $\mathbf{U}_k$ is independent of all other randomness and $\mathbb{E}[e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}] = 0$, conditioning on everything except for $\mathbf{U}_k$ gives

$$\mathbb{E}\Big[L(\boldsymbol{\theta}_{k-1} + \mathbf{U}'_k, \mathbf{X}_k, Y_k)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big)\Big] = \mathbb{E}\Big[L(\boldsymbol{\theta}_{k-1} + \mathbf{U}'_k, \mathbf{X}_k, Y_k)\mathbb{E}\big[e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big]\Big] = 0.$$

Thus, as in the original version, the anticipated loss vanishes in expectation and Theorem 1 in [26] holds without any changes. While the connection to gradient descent is appealing, the main obstacle for fast convergence of zeroth-order methods is the high variance.

2.2 The query model

A natural first step of a statistical analysis of biologically inspired learning is to study the properties of the zeroth-order update rule (2) for standard statistical models. The overall aim is to identify relevant models where this rule achieves the optimal convergence rates and/or outperforms other standard methods.

A fundamental problem in statistics and machine learning is the case where we want to learn a regression/feature vector $\boldsymbol{\theta}^\star$ such that for an independent draw $(\mathbf{X}, Y)$ with the same distribution as the (training) data $(\mathbf{X}_k, Y_k)$, we have $Y \approx \mathbf{X}^\intercal \boldsymbol{\theta}^\star$. Considering squared loss

$$L(\boldsymbol{\theta}, \mathbf{X}, Y) = \big(Y - \mathbf{X}^\intercal \boldsymbol{\theta}\big)^2,$$

the update formula (2) can then be rewritten as

$$\boldsymbol{\theta}_k \;=\; \boldsymbol{\theta}_{k-1} + \alpha_k \left(\big(Y_k - \mathbf{X}_k^\intercal(\boldsymbol{\theta}_{k-1} + \mathbf{U}_k)\big)^2 - \big(Y_k - \mathbf{X}_k^\intercal(\boldsymbol{\theta}_{k-1} + \mathbf{U}'_k)\big)^2 \right) \big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big). \qquad (3)$$

Interestingly, this formula does not require full knowledge of the covariate vector $\mathbf{X}_k$ and can be computed from queries. Here a query is defined as follows. Based on earlier seen queried observations and possibly extra randomness, the statistician chooses in the k-th round a query vector $\mathbf{v}_k$. Instead of the full k-th training sample $(\mathbf{X}_k, Y_k)$, one can only observe

$$Z_k = Y_k - \mathbf{X}_k^\intercal \mathbf{v}_k.$$

To realize (3), one needs to query each $(\mathbf{X}_k, Y_k)$ twice. This means that in the k-th round, we choose two query vectors $\mathbf{v}_k, \mathbf{v}'_k$ based on previously seen query values and extra randomness. As observations, we get

$$Z_k = Y_k - \mathbf{X}_k^\intercal \mathbf{v}_k, \quad \text{and} \quad Z'_k = Y_k - \mathbf{X}_k^\intercal \mathbf{v}'_k. \qquad (4)$$

To see that this observational scheme is indeed enough to compute the updates (3), we can argue by induction, assuming that the initialization $\boldsymbol{\theta}_0$ has been chosen independently of the data. The induction assumption is then, that $\boldsymbol{\theta}_k$ only depends on the queries $(Z_1, Z'_1, \ldots, Z_k, Z'_k)$ and exogenous randomness. Then, $\mathbf{v}_{k+1} = \boldsymbol{\theta}_k + \mathbf{U}_k$ and $\mathbf{v}'_{k+1} = \boldsymbol{\theta}_k + \mathbf{U}'_k$ are admissible query vectors with corresponding queries $Z_{k+1} = Y_{k+1} - \mathbf{X}_{k+1}^\intercal(\boldsymbol{\theta}_k + \mathbf{U}_{k+1})$ and $Z'_{k+1} = Y_{k+1} - \mathbf{X}_{k+1}^\intercal(\boldsymbol{\theta}_k + \mathbf{U}'_{k+1})$. The update formula implies that $\boldsymbol{\theta}_{k+1}$ only depends on the queries $(Z_1, Z'_1, \ldots, Z_{k+1}, Z'_{k+1})$ and exogenous randomness, completing the induction step.

Queries can be thought of as an attention mechanism that tell us where to look next to extract useful information about the data. A query model seems appropriate if the full input vector cannot be processed because, for instance, the data arrive too quickly. The information from the queries is insufficient to compute gradients. Hence, gradient descent methods cannot be run in this case.

There is a wide variety of previously considered query models in the statistical literature. Examples are querying large graphs to avoid storage problems [21] or learning causal relationships from path queries [7].

To finish the description of the model, we have to specify the distribution of the latent/unobserved variables $(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots$ We assume that those are i.i.d. and generated from the linear regression model with random design, that is, $\mathbf{X}_1, \mathbf{X}_2, \dots$ are drawn i.i.d. from some unknown distribution $P_{\mathbf{X}}$ and the response variables are

$$Y_i = \mathbf{X}_i^\intercal \boldsymbol{\theta}^\star + \varepsilon_i, \quad \text{for } \ i = 1, \dots \tag{5}$$

for i.i.d. noise variables $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ that are also independent of the covariates $\mathbf{X}_1, \dots$ We assume that $\sigma > 0$ is known.

Let $(\mathbf{X}, Y)$ be a new and independent sample with the same distribution as the training samples. Write

$$Q := \mathbb{E}[\mathbf{X}\mathbf{X}^\intercal]$$

for the (uncentered) covariance matrix of the design vectors. Consider a possibly randomized estimator $\widehat{\boldsymbol{\theta}}$. By construction, $\widehat{\boldsymbol{\theta}}$ is independent of a test point $(\mathbf{X}, Y)$. Rewriting $Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\theta}} = (Y - \mathbf{X}^\intercal \boldsymbol{\theta}^\star) + \mathbf{X}^\intercal(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)$, conditioning on $\widehat{\boldsymbol{\theta}}$ and using tower rule, the excess risk of $\widehat{\boldsymbol{\theta}}$ is

$$\begin{aligned}
\mathbb{E}\left[(Y - \mathbf{X}^\intercal \widehat{\boldsymbol{\theta}})^2 - (Y - \mathbf{X}^\intercal \boldsymbol{\theta}^\star)^2\right] &= \mathbb{E}\left[2(Y - \mathbf{X}^\intercal \boldsymbol{\theta}^\star)\mathbf{X}^\intercal(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star) + (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\intercal \mathbf{X}\mathbf{X}^\intercal(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \mathbb{E}\left[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)^\intercal Q(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star)\right] \\
&= \operatorname{tr}(QS)
\end{aligned} \tag{6}$$

with $S \coloneqq \mathbb{E}\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right)\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\right)^\intercal\right]$.

[9] relates gradient descent with dropout to a vector autoregressive (VAR) process with random coefficients [24]. A similar argument can be made here. Defining $\mathbf{W}_k := \boldsymbol{\theta}_k - \boldsymbol{\theta}^\star$, $G_k := I - 2\alpha_k\left(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\right)\left(\mathbf{U}_k' - \mathbf{U}_k\right)^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal$ and $\boldsymbol{\xi}_k := 2\alpha_k \varepsilon_k \mathbf{X}_k^\intercal(\mathbf{U}_k' - \mathbf{U}_k)\left(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\right) + \alpha_k\left((\mathbf{X}_k^\intercal \mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal \mathbf{U}_k')^2\right)\left(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\right)$, the update formula can be written in the form of a lag one VAR process

$$\mathbf{W}_k = G_k \mathbf{W}_{k-1} + \boldsymbol{\xi}_k, \tag{7}$$

with independent random coefficients G_k and independent noise/innovation variables $\boldsymbol{\xi}_k$. It can be checked that the noise is centered, that is,

$$\mathbb{E}[\boldsymbol{\xi}_k] = 0. \tag{8}$$

A proof of this fact and a derivation of (7) can be found in Appendix A. As we can tune the learning rate α_k and consider the parameter A from the uniform distribution as hyperparameter, it is interesting to work out the scaling of the different terms in these parameters. For small A, $e^{-\mathbf{U}_k} - e^{\mathbf{U}_k} \approx -2\mathbf{U}_k$ which is of order A. This means that the term $2\alpha_k\left(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\right)\left(\mathbf{U}_k' - \mathbf{U}_k\right)^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal$ in the definition of G_k will be of order $\alpha_k A^2$. The same is true for the first term of $\boldsymbol{\xi}_k$ while the second term of $\boldsymbol{\xi}_k$ scales like $O(\alpha_k A^3)$. This suggests to define

$$\textit{effective learning rate} := \alpha_k A^2. \tag{9}$$

This quantity will remain constant if A is decreased by a factor $\gamma < 1$, and, concurrently, α_k is increased by a factor γ^{-2}. Thus by making γ and thus A small, the term of the order $O(\alpha_k A^3)$ becomes negligible. While the considered optimization scheme is motivated by Hebbian learning with A a small constant, treating A as a hyperparameter also suggests to choose a small A.

The updates (3) are independent of $Q = \mathbb{E}[\mathbf{X}\mathbf{X}^\intercal]$. As common in the literature on gradient descent, the chosen learning rate will, however, depend on the smallest eigenvalue of Q.

Applying (6) to the estimator $\boldsymbol{\theta}_k$, we need to control

$$S_k := \mathbb{E}[(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)^\intercal] = \mathbb{E}[\mathbf{W}_k \mathbf{W}_k^\intercal].$$

We can relate this to the previous iterate via $\mathbf{W}_k \mathbf{W}_k^\intercal = G_k \mathbf{W}_{k-1} \mathbf{W}_{k-1}^\intercal G_k + \boldsymbol{\xi}_k \boldsymbol{\xi}_k^\intercal + G_k \mathbf{W}_{k-1} \boldsymbol{\xi}_k^\intercal + \boldsymbol{\xi}_k \mathbf{W}_{k-1}^\intercal G_k$. Based on this identity, we will now derive a recursive formula for S_k. Assume that $\mathbf{U}, \mathbf{U}' \sim \mathrm{Unif}[-A, A]^d$ are independent and also independent of $\mathbf{X}$. Set $\mathbf{D} = e^{-\mathbf{U}} - e^{\mathbf{U}}$ and let $\mathbf{x}$ be a fixed d-dimensional vector. The following two matrices characterize the interaction between the d-dimensional noise vectors $(\mathbf{U}, \mathbf{U}')$,

$$\begin{aligned} V(\mathbf{x}) &\coloneqq \mathbb{E}\big[\mathbf{x}^\intercal (\mathbf{U}' - \mathbf{U}) \mathbf{D}\mathbf{D}^\intercal (\mathbf{U}' - \mathbf{U})^\intercal \mathbf{x}\big] \\ W &\coloneqq \mathbb{E}\big[\left((\mathbf{X}^\intercal \mathbf{U})^2 - (\mathbf{X}^\intercal \mathbf{U}')^2\right)^2 \mathbf{D}\mathbf{D}^\intercal\big]. \end{aligned} \tag{10}$$

Lemma 1. *(i) the interaction terms $G_k \mathbf{W}_{k-1} \boldsymbol{\xi}_k^\intercal$ and $\boldsymbol{\xi}_k \mathbf{W}_{k-1}^\intercal G_k$ have mean zero,*

(ii) $\mathrm{Cov}(\boldsymbol{\xi}_k) = \mathbb{E}[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\intercal] = 4\alpha_k^2 \sigma^2 \,\mathbb{E}\,[V(\mathbf{X})] + \alpha_k^2 W$,

(iii) if $\mu := -\mathbb{E}[U(e^{-U} - e^U)]$ with $U \sim \mathrm{Unif}([-A, A])$, then,

$$\begin{aligned} S_k = \big(I - 2\alpha_k \mu Q\big) S_{k-1} \big(I - 2\alpha_k \mu Q\big) + 4\alpha_k^2 \left\{\mathbb{E}\left[\mathbf{X}^\intercal S_{k-1} \mathbf{X} V(\mathbf{X})\right] - \mu^2 Q S_{k-1} Q\right\} \\ + 4\alpha_k^2 \sigma^2 \,\mathbb{E}\,[V(\mathbf{X})] + \alpha_k^2 W. \end{aligned}$$

and moreover,

$$S_k \leq \|S_{k-1}\| \Big\| I - 4\alpha_k \mu Q + 4\alpha_k^2 \,\mathbb{E}\left[\|\mathbf{X}\|^2 V(\mathbf{X})\right] \Big\| + 4\alpha_k^2 \sigma^2 \,\mathbb{E}\,[V(\mathbf{X})] + \alpha_k^2 W. \tag{11}$$

In summary, S_k is an affine function in S_{k-1}.

As discussed before, the last term $\alpha_k^2 W$ can be made negligible by choosing a sufficiently small parameter A in the uniform distribution. The term μ is $O(A^2)$ and, under moment assumptions on $\mathbf{X}$, the respective order of the terms $\| \mathbb{E}\,[V(\mathbf{X})] \|$ and $\| \mathbb{E}\left[\|\mathbf{X}\|^2 V(\mathbf{X})\right] \|$ is $O(A^4 d)$ and $O(A^4 d^2)$. To decrease the norm of the error matrix S_k in (11), the factor $I - 4\alpha_k \mu Q + 4\alpha_k^2 \,\mathbb{E}\left[\|\mathbf{X}\|^2 V(\mathbf{X})\right]$ has to remain below the identity matrix. This implies that $\alpha_k A^2 Q \gtrsim \alpha_k^2 A^4 d^2 I$ and motivates to choose the effective learning rate $\alpha_k A^2$ in (9) to be bounded by $\lesssim \lambda_{\min}(Q)/d^2$. Compared to standard choices such as $1/k$, the learning rate in the beginning is thus small and learning makes little progress.

If $k \gtrsim d^2$, effective learning rate $1/k$ is possible by choosing $\alpha_k A^2 \asymp 1/(k \vee \lambda_{\min}(Q) d^2)$. The rate is then determined by the contributions $4\alpha_k^2 \sigma^2 \| \mathbb{E}\,[V(\mathbf{X})] \| = O(\sigma^2 d/k^2)$. The $1/k^2$ becomes a $1/k$ as the contributions from the previous $O(k)$ iterates add up. Up to a $\log^2(d)$-factor, this explains the convergence rate stated in (13) below.

To simplify the exposition, we work now with specific choices for the parameter A in the uniform distribution and the learning rate α_k. The expressions also depend on the variance of the noise σ^2 in the linear regression model, which is assumed to be known. To obtain a convenient expressions for the upper bound, we moreover introduce

$$M_k := \max_{i=1,\dots,d} 1 \vee E\big[X_i^k\big],$$

where X_i denotes the i-th component of $\mathbf{X}$.

Theorem 1. *Assume $d \geq 9$. Let $A = \sigma/\sqrt{d}$, and choose the learning rate*

$$\alpha_k = \frac{11B\log(d)}{A^2\lambda_{\min}(Q)(Bk + d^2\log(d))} \quad \textit{with} \quad B := 1 \wedge \frac{\lambda_{\min}(Q)^2}{2904M_4}. \tag{12}$$

Then there exists a constant $C = C(M_4, \lambda_{\min}(Q))$, such that for all $k > 2d^2\log(d)/B$, we have

$$\big\|\mathbb{E}[(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)^\intercal]\big\| \leq \Big(\frac{2d^2\log(d)}{Bk}\Big)^{\log(d)}\|S_0\| + C\frac{\sigma^2 d\log^2(d)}{k}. \tag{13}$$

If moreover $\operatorname{tr}(Q) \leq d$, the excess risk (6) is bounded by

$$\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)^\intercal Q(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] \leq d\Big(\frac{2d^2\log(d)}{Bk}\Big)^{\log(d)}\|S_0\| + C\frac{\sigma^2 d^2\log^2(d)}{k}.$$

Assuming $\operatorname{tr}(Q) \leq d$ is natural and includes in particular the case that $Q = \mathbb{E}[\mathbf{X}\mathbf{X}^\intercal]$ is the identity matrix.

Initializing $\boldsymbol{\theta}_0 = \mathbf{0}$ with the zero vector, we have $S_0 = \boldsymbol{\theta}^\star(\boldsymbol{\theta}^\star)^\intercal$ and with Lemma 4, $\|S_0\| = \|\boldsymbol{\theta}^\star\|^2$.

The rate consists of two terms. For any $\kappa > 0, d > e^\kappa$, and $k \geq 2e^{\gamma+3\kappa}d^2\log(d)/B$, we have

$$\Big(\frac{2d^2\log(d)}{Bk}\Big)^{\log(d)} \lesssim \frac{1}{d^\gamma k^\kappa}. \tag{14}$$

A proof of this inequality is given in Appendix A.4. The inequality states that for sufficiently large $k \gtrsim d^2\log(d)$, the first term is negligible and the rate is $\sigma^2 d^2\log^2(d)/k$. This rate is slower than the usual rate $\sigma^2 d/k$ if the full data are observed in the linear regression model. The $\log^2(d)$ factor seems to be an artifact of the proof and we will later derive nearly-matching lower bounds.

Why do we loose a factor d due to querying? It is tempting to assume that observed random variables from the query model and the linear regression model are equally informative. Then one can link the additional factor d for the query model to the loss of information in the data: While in the linear regression model we observe in every round a d-dimensional covariate vector $\mathbf{X}_k$ together with a real-valued response Y_k, the query model observes in each round two real-valued queries. The total number of observed random variables in the query model is thus decreased by an order $O(d)$. Differently speaking the query model needs $O(d)$ more iterations to receive the same number of observed variables. The concepts in [28] might be suitable to formalize such an argument.

The result is similar in spirit to the bounds obtained in [8] for another biologically motivated learning rule called (weight-perturbed) forward gradient descent [6, 25]. For a sequence of i.i.d.

random vectors $\boldsymbol{\zeta}_1, \boldsymbol{\zeta}_2, \ldots \mathcal{N}(0, I_d)$, forward gradient descent with learning rate α'_k is given by the update rule

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} - \alpha'_k \big(\nabla L(\boldsymbol{\theta}_{k-1})\big)^\intercal \boldsymbol{\zeta}_k \boldsymbol{\zeta}_k, \quad k = 1, 2, \ldots$$

The similarity is best explained for squared loss L. In this case, $\nabla L(\boldsymbol{\theta}) = -2\big(Y_k - \mathbf{X}_k^\intercal \boldsymbol{\theta}\big)\mathbf{X}_k$. If instead of an independent draw, we set $\mathbf{U}'_k := -\mathbf{U}_k$, then the optimization scheme (3) becomes

$$\begin{aligned}\boldsymbol{\theta}_k &= \boldsymbol{\theta}_{k-1} + \alpha_k \left(\big(Y_k - \mathbf{X}_k^\intercal(\boldsymbol{\theta}_{k-1} + \mathbf{U}_k)\big)^2 - \big(Y_k - \mathbf{X}_k^\intercal(\boldsymbol{\theta}_{k-1} - \mathbf{U}_k)\big)^2 \right) \big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big) \\ &= \boldsymbol{\theta}_{k-1} + 2\alpha_k \nabla L(\boldsymbol{\theta}_{k-1})^\intercal \mathbf{U}_k \big(e^{\mathbf{U}_k} - e^{-\mathbf{U}_k}\big).\end{aligned}$$

For small A, $e^{\mathbf{U}_k} - e^{-\mathbf{U}_k} \approx 2\mathbf{U}_k$, which means that $\boldsymbol{\theta}_k \approx \boldsymbol{\theta}_{k-1} + 4\alpha_k \nabla L(\boldsymbol{\theta}_{k-1})^\intercal \mathbf{U}_k \mathbf{U}_k$. Whereas forward gradient descent has been proposed for normally distributed $\boldsymbol{\zeta}_k$, the gradient-free optimization scheme here is closely related to the case where these vectors are sampled from a uniform distribution. The choice $\mathbf{U}'_k := -\mathbf{U}_k$ leads to less remainder terms in the analysis and less additional noise, but we find this choice less appealing to model Hebbian learning in the brain.

3 Lower bound for adaptive queries

We now derive nearly-matching lower bounds. To make the lower bound rigorous, we need to formalize the query model. The observed data consists of $2k$ query vectors $\mathbf{v}_1, \mathbf{v}'_1, \ldots, \mathbf{v}_k, \mathbf{v}'_k$ and a random vector $(Z_1, Z'_1, \ldots, Z_k, Z'_k)$ of length $2k$.

The unobserved/latent i.i.d. pairs $(\mathbf{X}_\ell, Y_\ell)$, $\ell = 1, \ldots, k$ are generated from the linear regression model (5). For the lower bounds we will moreover assume that the design distribution $P_{\mathbf{X}}$ is $\mathcal{N}(0, I_d)$ with I_d the $d \times d$ identity matrix. This implies that $Q = E[\mathbf{X}\mathbf{X}^\intercal] = I_d$.

Let $(\mathcal{G}_\ell)_{\ell \geq 0}$ be a filtration generated by exogenous randomness. We assume that for any $\ell = 1, 2, \ldots$, the query vectors $\mathbf{v}_\ell, \mathbf{v}'_\ell$ are measurable with respect to the σ-algebra

$$\mathcal{F}_{\ell-1} := \sigma\big(Z_1, Z'_1, \ldots, Z_{\ell-1}, Z'_{\ell-1}, \mathbf{v}_1, \mathbf{v}'_1, \ldots, \mathbf{v}_{\ell-1}, \mathbf{v}'_{\ell-1}\big) \times \mathcal{G}_{\ell-1}.$$

This means that the query vectors might depend on past queries, past query vectors and exogenous randomness. Given $\mathbf{v}_\ell, \mathbf{v}'_\ell$, we observe in the ℓ-th step the queries (Z_ℓ, Z'_ℓ) with

$$Z_\ell = Y_\ell - \mathbf{X}_\ell^\intercal \mathbf{v}_\ell, \qquad Z'_\ell = Y_\ell - \mathbf{X}_\ell^\intercal \mathbf{v}'_\ell. \tag{15}$$

Finally, an estimator $\widehat{\boldsymbol{\theta}}_k$ is any measurable function in $\mathcal{F}_k$. This means that an estimator is allowed to depend on all observed queries, all observed query vectors and the exogenous randomness.

To allow for two queries instead of one makes the lower bounds considerable more involved as Z_ℓ and Z'_ℓ are dependent by virtue of sharing $\mathbf{X}_\ell, Y_\ell$. One query is, however, not enough to realize the zeroth-order scheme (3).

Before stating the lower bounds, we derive an equivalent representation of this model. Two statistical models $(P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ and $(Q_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta)$ with the same parameter space Θ are equivalent if there exist Markov kernels M, M' that are independent of unknown parameters such that $Q_{\boldsymbol{\theta}} = M P_{\boldsymbol{\theta}}$ and $P_{\boldsymbol{\theta}} = M' Q_{\boldsymbol{\theta}}$ for all $\boldsymbol{\theta} \in \Theta$.

Lemma 2. *Consider the query model* $\{Z_1, Z_1', \ldots, Z_k, Z_k', \mathbf{v}_1, \mathbf{v}_1', \ldots, \mathbf{v}_k, \mathbf{v}_k'\}$ *with* (15) *replaced by*

$$Z_\ell = Y_\ell - \mathbf{X}_\ell^\intercal \mathbf{v}_\ell, \qquad Z_\ell' = \mathbf{X}_\ell^\intercal \mathbf{v}_\ell', \qquad\qquad \ell = 1, 2, \ldots \tag{16}$$

Both models are statistically equivalent.

For the lower bounds we will work in the transformed query model (16). If we choose the query vector $\mathbf{v}_\ell' = 0$, then $Z_\ell' = 0$ and this is as informative as only querying the model once.

For the model with queries (16), we have

$$\begin{pmatrix} Z_\ell \\ Z_\ell' \end{pmatrix} \Big| \, \mathbf{v}_\ell, \mathbf{v}_\ell' \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell\|^2 & \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}_\ell' \rangle \\ \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}_\ell' \rangle & \|\mathbf{v}_\ell'\|^2 \end{pmatrix} \right). \tag{17}$$

using that, we also assumed $\mathbf{X}_k \sim \mathcal{N}(0, I)$ and thus $Q = I_d$.

The choice of the query vectors up to round k, will be called the *query strategy* (up to round k) and is denoted by $\mathcal{V}_k$. It is important to recall that the query strategy cannot depend on $\boldsymbol{\theta}$ as it is unknown.

The data distribution of $(Z_1, Z_1', \ldots, Z_k, Z_k', \mathbf{v}_1, \mathbf{v}_1', \ldots, \mathbf{v}_k, \mathbf{v}_k')$ will be denoted by $P_{\boldsymbol{\theta}, \mathcal{V}_k}$ and depends on the underlying parameter $\boldsymbol{\theta}$ and the query strategy $\mathcal{V}_k$. Since the choice of the query strategy $\mathcal{V}_k$ is part of the estimation procedure, the right notion of the minimax estimation risk is

$$\inf_{\widehat{\boldsymbol{\theta}},\ \mathcal{V}_k} \sup_{\boldsymbol{\theta} \in \Theta} \mathbb{E}_{\boldsymbol{\theta}, \mathcal{V}_k} \left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right],$$

where the infimum is taken over all query strategies and all estimators. The parameter space Θ will be chosen as the Euclidean ball with radius R, that is,

$$B_R(0) := \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\| \le R\}.$$

Theorem 2. *If* $d \ge 3$ *and* $k \ge d^2$*, then,*

$$\inf_{\widehat{\boldsymbol{\theta}},\ \mathcal{V}_k} \sup_{\boldsymbol{\theta} \in B_R(0)} \mathbb{E}_{\boldsymbol{\theta}, \mathcal{V}_k} \left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right] \ge \frac{1}{162} \Big(1 - \frac{1}{\sqrt{2}}\Big) \Big(R^2 \wedge \frac{d^2}{k} \sigma^2 \Big).$$

The statement assumes that $k \ge d^2$. A similar condition ($k > 2d^2 \log(d)/B$) appears also in the corresponding upper bound in Theorem 1.

Interestingly, the rate does not depend on the radius R for large k. This indicates that the statistical main difficulty is to recover the direction of the true regression vector.

The same upper and lower bounds hold if instead of (4), we observe the squared queries $(Z_1^2, (Z_1')^2, \ldots, Z_k^2, (Z_k')^2)$. Indeed the squares are sufficient to implement the biologically inspired updating rule (3). Since $(Z_1^2, (Z_1')^2, \ldots, Z_k^2, (Z_k')^2)$ is at most as informative as $(Z_1, Z_1', \ldots, Z_k, Z_k')$, the lower bounds remain true.

4 Minimax risk for nonadaptive queries

What is the advantage to use previous information to select the next query vector? To answer this, we now consider query vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$ that are chosen before the data are revealed. Denote by $\mathcal{M}_k$ the space of all such query strategies. For convenience, we will moreover assume throughout this section that the design distribution is standard multivariate normal, this means that

$$\mathbf{X}_1, \mathbf{X}_2, \dots \sim P_{\mathbf{X}} = \mathcal{N}(0, I_d), \quad \text{i.i.d.} \tag{18}$$

We show in this section that the minimax estimation risk with non-adaptive query strategies in $\mathcal{M}_k$ and parameter space Θ the Euclidean ball $B_R(0)$ is

$$\inf_{\widehat{\boldsymbol{\theta}},\ \mathcal{V}_k \in \mathcal{M}_k} \sup_{\boldsymbol{\theta} \in B_R(0)} \mathbb{E}_{\boldsymbol{\theta}, \mathcal{V}_k} \left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right] \asymp R^2 \wedge \frac{d^2}{k} (R \vee \sigma)^2. \tag{19}$$

For the upper bound we construct the following estimator. If $k \le 2d^2((\sigma^2/R^2) \vee 1)$, take $\widehat{\boldsymbol{\theta}} = 0$. If $k > 2d^2((\sigma^2/R^2) \vee 1)$, we have $d(k/(2d) + 1) = k/2 + d \le k$ and can therefore partition the index set $\{1, \dots, k\}$ into d blocks $\mathcal{B}_1, \dots, \mathcal{B}_d$ such that each block has cardinality $\ge k/(2d)$. Set $\mathbf{v}_j = (R \vee \sigma)\mathbf{e}_s$ if $j \in B_s$ with $\mathbf{e}_s$ the s-th standard basis vector. Thanks to (18), the data are then given by

$$Z_j \sim \mathcal{N}\big(0, \sigma^2 + \|\boldsymbol{\theta} - (R \vee \sigma)\mathbf{e}_s\|^2\big) = \mathcal{N}\Big(0, \sigma^2 + \|\boldsymbol{\theta}\|^2 - 2(R \vee \sigma)\boldsymbol{\theta}_s + (R \vee \sigma)^2\Big), \quad \text{if } \ j \in B_s$$

and the estimator for the s-th component of $\boldsymbol{\theta}$ is in this case

$$\widehat{\theta}_s := \frac{\sigma^2 + \|\boldsymbol{\theta}\|^2 + (R \vee \sigma)^2 - |\mathcal{B}_s|^{-1} \sum_{r \in \mathcal{B}_s} Z_r^2}{2(R \vee \sigma)}, \quad \text{for } s = 1, \dots, d, \tag{20}$$

with $|\mathcal{B}_s|$ the cardinality of the set $\mathcal{B}_s$. Whenever $r \in \mathcal{B}_s$, we have $\mathbb{E}[Z_r^2] = \sigma^2 + \|\boldsymbol{\theta}\|^2 - 2(R \vee \sigma)\boldsymbol{\theta}_s + (R \vee \sigma)^2$, implying that

$$\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \dots, \widehat{\theta}_d)$$

is an unbiased estimator for $\boldsymbol{\theta}$.

Theorem 3. *Assume* (18). *For the estimator* $\widehat{\boldsymbol{\theta}}$ *defined componentwise in* (20), *we have*

$$\sup_{\boldsymbol{\theta} \in B_R(0)} \mathbb{E}_{\boldsymbol{\theta}, \mathcal{V}_k} \left[\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta} \right\|^2 \right] \le 25 \left(R^2 \wedge \frac{d^2}{k} (R \vee \sigma)^2 \right).$$

The estimator uses knowledge of R and σ^2. If these quantities are unknown, estimation seems unequal harder, in particular if R is small.

The query vectors $\mathbf{v}_j$ can be thought of as test functions or features for $\boldsymbol{\theta}$. In the construction of the estimator, they have norm $\|\mathbf{v}_j\| = (R \vee \sigma)$. Interestingly, if $\sigma > R$, the norm exceeds R and the $\mathbf{v}_j$ are themselves not in the parameter space $\Theta = B_R(0)$.

Proof. For $k \le 2d^2((\sigma^2/R^2) \vee 1)$, $\widehat{\boldsymbol{\theta}} = 0$ and the result follows since $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 = \|\boldsymbol{\theta}\|^2 \le R^2$.

It remains to show that for $k > 2d^2((\sigma^2/R^2) \vee 1)$, the rate is bounded by $25\frac{d^2}{k}(R \vee \sigma)^2$. The bias-variance decomposition yields

$$\mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\left[\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right\|^2\right] = \sum_{s=1}^{d} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\left[\left(\widehat{\boldsymbol{\theta}}_s - \boldsymbol{\theta}_s\right)^2\right] = \sum_{s=1}^{d} \mathrm{Bias}^2_{\boldsymbol{\theta},\mathcal{V}_k}\left(\widehat{\boldsymbol{\theta}}_s\right) + \sum_{s=1}^{d} \mathrm{Var}_{\boldsymbol{\theta},\mathcal{V}_k}\left(\widehat{\boldsymbol{\theta}}_s\right).$$

As we have already shown that $\widehat{\boldsymbol{\theta}}$ is unbiased, it remains to bound the variances. For $\xi \sim \mathcal{N}(0, a^2)$, we have $\mathrm{Var}(\xi^2) = \mathbb{E}[\xi^4] - \mathbb{E}^2[\xi^2] = 3a^4 - a^2 = 2a^2$. Using the definition of the estimator in (20), the independence of $Z_1, \dots, Z_k$, that $\|\boldsymbol{\theta}\|^2 \le R^2$, that $\sigma^2 + \|\boldsymbol{\theta}\|^2 - 2R\boldsymbol{\theta}_s + R^2 \le 5(R \vee \sigma)^2$, and that by construction of the blocks $|\mathcal{B}_s| \ge k/(2d)$, we obtain for any $j \in \mathcal{B}_s$,

$$\mathrm{Var}_{\boldsymbol{\theta},\mathcal{V}_k}\left(\widehat{\boldsymbol{\theta}}_s\right) = \frac{\mathrm{Var}(Z_j^2)}{(2(R \vee \sigma))^2 |\mathcal{B}_s|} = \frac{2(\sigma^2 + \|\boldsymbol{\theta}\|^2 - 2R\boldsymbol{\theta}_s + R^2)^2}{4(R \vee \sigma)^2 |\mathcal{B}_s|} \le \frac{50(R \vee \sigma)^4}{4(R \vee \sigma)^2 |\mathcal{B}_s|} \le \frac{25d}{k}(R \vee \sigma)^2.$$

Summing over $s = 1, \dots, d$ gives another factor d and thus the claim follows. □

We now state the corresponding lower bound.

Theorem 4. *Assume* (18). *If* $d \ge 6$, *then for any* $k = 1, 2, \dots$

$$\inf_{\widehat{\boldsymbol{\theta}},\ \mathcal{V}_k \in \mathcal{M}_k} \sup_{\boldsymbol{\theta} \in B_R(0)} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\right] \ge 2^{-18}\left(R^2 \wedge \frac{d^2}{k}(R \vee \sigma)^2\right).$$

Together with Theorem 3, this shows (19). While in the adaptive setting, we had to impose the restriction $k \gtrsim d^2 \log(d)$ for the upper bound and $k \ge d^2$ for the lower bound, the derived rate in the non-adaptive setting holds for all sample sizes k.

Compared to the adaptive case, proving the lower bound in the non-adaptive case is considerably more involved. Reasons are that in this case, one additional regime occurs in the rate. To deal with this regime requires to show that whatever the choice of the query vectors is, one can find a regression vector $\boldsymbol{\theta}^\star$ in the parameter space that is far away. The fact that we allow for two queries per sample $(\mathbf{X}_\ell, Y_\ell)$ makes that 'far away' has to be interpreted with respect to some tube that is generated by the pair of query vectors $(\mathbf{v}_\ell, \mathbf{v}'_\ell)$.

We believe that the constant 2^{-18} in the lower bound can be improved significantly at the expense of a more technical proof.

The upper bound only needs one query per iteration and the lower bound is derived for two queries per iteration. This already proves that if we can only query once in every iteration, the minimax rate remains the same.

Based on the derived lower bound, we can now quantify the gap between adaptive and non-adaptive design. A natural setting is to allow that all parameters are of order one. This means that $\|\boldsymbol{\theta}\|^2 = \sum_{j=1}^{d} \theta_j^2$ is of order $O(d)$ and motivates to choose $R = \sqrt{d}$.

Corollary 1. *Assume* (18). *If* $\sigma = 1$, $Q = I, \boldsymbol{\theta}_0 = 0$, *and* $R = \sqrt{d}$, *then for all* $k \ge 2e^5 d^2 \log(d)/B$ *(with* B *the constant in* (12) *) and* $d \ge 8$, *the upper for the adaptive design yields the convergence rate*

$$\frac{d^2 \log^2(d)}{k}$$

while the minimax rate for the non-adaptive design is

$$\frac{d^3}{k}.$$

The result implies that constructing queries based on previously seen data improves the rate by a factor d^{-1} (up to logarithms). The improvement will become even more pronounced if R increases. Indeed, the upper bound in the adaptive query setting will remain $d^2 \log^2(d)/k$, while the minimax rate for the non-adaptive query setting becomes d^2R^2/k. We do not have a convincing heuristic argument explaining the gap in the rates. However, it is clear that the adaptive query setting can learn over time about the direction of the true $\boldsymbol{\theta}^*$, whereas in the non-adaptive query setting one has to spread out the query vectors equally over all possible directions. This is also clearly visible in the construction of the estimator in (20).

For the related problem of adaptive sensing, it has been found in [2] that adaptation improves the rate by at most log-factors. In this setting there are no queries and one can choose in the i-th iteration a design vector $\mathbf{X}_i$ based on past observations and will then observe $Y_i = \mathbf{X}_i^\intercal \boldsymbol{\theta}^\star + \varepsilon_i$ with independent $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. If s is the number of non-zero components of $\boldsymbol{\theta}^\star$, it is shown that the risk $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}^\star\|^2]$ of any estimator $\widehat{\boldsymbol{\theta}}$ based on k measurements is lower bounded by $\gtrsim \sigma^2 s/k$. On the contrary, in the non-adaptive setting with $\mathbf{X}_i$ chosen i.i.d. and independent of previous data, the Dantzig selector $\widehat{\boldsymbol{\theta}}_k^D$ based on $k \gtrsim s\log(d/s)$ measurements achieves estimation rate $\mathbb{E}[\|\widehat{\boldsymbol{\theta}}_k^D - \boldsymbol{\theta}^\star\|^2] \lesssim \sigma^2 s \log(d)/k$. Thus the gain in the convergence rate of an adaptive sampling strategy is here at most a factor $\log(d)$ in the rate. The reason why adaptation hardly improves the rate in this setting is attributed in [2] to the difficulty to recover the support of the sparse regression vector $\boldsymbol{\theta}^\star$.

5 Related literature

In zeroth-order stochastic convex optimization the task is to learn a minimizer of an unknown convex function $f(\boldsymbol{\theta}) = \mathbb{E}_\Xi[f(\boldsymbol{\theta}, \Xi)]$. The vanilla framework is to sequentially issue queries $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots$ and receive noisy observations $f(\boldsymbol{\theta}_k, \Xi_k)$ for i.i.d. unobserved $\Xi_1, \Xi_2, \ldots$. The particular case of linear regression $Y = \mathbf{X}^\intercal \boldsymbol{\theta}^* + \epsilon$ corresponds to $\Xi = (\mathbf{X}, \epsilon)$ with d-dimensional covariate vectors $\mathbf{X}$ drawn from an unknown distribution, noise ϵ, and feedback $f(\boldsymbol{\theta}, (\mathbf{X}, \epsilon)) = (\mathbf{X}^\intercal(\boldsymbol{\theta}^\star - \boldsymbol{\theta}) + \epsilon)^2 = (Y - \mathbf{X}^\intercal\boldsymbol{\theta})^2$. In the literature (we refer to [19, 5] for excellent surveys), rates for algorithms and lower bounds are organised based on three main distinctions:

- in the *optimization* literature the objective is the gap $f(\widehat{\boldsymbol{\theta}}_k) - \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$ of a proposed evaluation point $\widehat{\boldsymbol{\theta}}_k$, while in the *bandit* literature the objective is the regret of the queries issued $\sum_{t=1}^k (f(\boldsymbol{\theta}_t) - \min_{\boldsymbol{\theta}} f(\boldsymbol{\theta}))$. The latter bounds the former by online to batch conversion, but there are interesting separations in minimax rates [27].

- in *one-point feedback* the learner issues a point $\boldsymbol{\theta}$ and observes $f(\boldsymbol{\theta}, \Xi)$, where the noise Ξ is i.i.d. between queries [13]. In *two-point feedback* the learner issues a pair of points $\boldsymbol{\theta}, \boldsymbol{\theta}'$ and observes $f(\boldsymbol{\theta}, \Xi)$ and $f(\boldsymbol{\theta}', \Xi)$ with shared noise Ξ. Different rates can occur between one and two-point feedback.

- in addition to convexity of f, one may get better rates by assuming that $f(\cdot)$ (or $f(\cdot,\xi)$ for every ξ) are Lipschitz, smooth, higher-order β-smooth [4], and/or strongly convex.

In [12], upper and lower bounds are derived for the gap $f(\widehat{\boldsymbol{\theta}}_k) - \min_{\boldsymbol{\theta}\in\Theta} f(\boldsymbol{\theta})$. Under the imposed conditions and ignoring the dependence on the number of parameters d, the gap decreases in k with the rate $1/\sqrt{k}$, which is slower than the $1/k$ rate obtained here. The lower bounds are obtained for linear $f(\boldsymbol{\theta}) = \mathbb{E}[\boldsymbol{\theta}^\intercal \mathbf{X}] = \boldsymbol{\theta}^\intercal \mathbb{E}[\mathbf{X}]$ with unknown distribution $\mathbf{X}$. In this case, the Hessian is zero and the minimizer will lie at the boundary of the parameter space Θ. Therefore the rates strongly depend on the choice of Θ.

In the case of linear regression, the function $f(\boldsymbol{\theta}) = E[(Y - \mathbf{X}^\intercal\boldsymbol{\theta})^2]$ is a convex quadratic defined on the entire Euclidean space $\mathbb{R}^d$, and as such strongly convex and smooth (of infinite order) but not Lipschitz. For one-point feedback, [22] proves that, under suitable assumptions, the averaged iterations $\overline{\boldsymbol{\theta}}_k = \frac{1}{k}\sum_{\ell=1}^k \boldsymbol{\theta}_\ell$ converge with rate $\frac{d^2}{k}(\frac{k}{d})^{1/\beta}$ for any $\beta > 0$.

Keeping the dimension d fixed and letting the number of iterations tend to infinity, [16] derives a CLT for the average over all iterates (Ruppert-Polyak average).

To complete this literature overview, we briefly mention related approaches. [27] considers a zeroth-order method to learn a minimizer of the function $F(\mathbf{w}) = \mathbf{w}^\intercal A\mathbf{w} + \mathbf{b}^\intercal\mathbf{w} + c$ for unknown $d \times d$ matrix A, d-dimensional vector $\mathbf{b}$ and scalar c. In every iteration, one can query the function F once. This is not a statistical task, as there are no data. Under suitable conditions, the considered zeroth-order method achieves the rate d^2/k, with k the number of iterations and it is moreover shown that this rate is optimal. Furthermore, optimal rates are also known in the case that f is strongly convex and smooth [1, 27].

These rates can further be contrasted to those for stochastic first-order feedback, where the feedback for a query θ is the stochastic gradient $\nabla f(\theta, \Xi)$. Here [3] show a gap rate of order $\frac{\sigma^2 d + \|\hat{\theta}_0 - \theta_*\|^2}{k}$ for the average iterate of SGD with constant step size. In [11] these rates are further improved with acceleration to order $\frac{\sigma^2 d}{k} + \frac{\|\hat{\theta}_0 - \theta_*\|^2}{k^2}$, matching lower bounds in both contributions. The results were extended beyond the least squares setting in [18] and the related problem of logistic regression was analyzed in the stochastic optimization [4] and bandit settings [15]. For logistic regression the function f is Lipschitz and higher-order smooth but not strongly convex.

A Proofs for the upper bound

Proof of (7)*:* Expanding the squares yields

$$\boldsymbol{\theta}_k \;=\; \boldsymbol{\theta}_{k-1} + \alpha_k\Big(2(Y_k - \mathbf{X}_k^\intercal\boldsymbol{\theta}_{k-1})\mathbf{X}_k^\intercal(\mathbf{U}_k' - \mathbf{U}_k) + (\mathbf{X}_k^\intercal\mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal\mathbf{U}_k')^2\Big)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big)$$

Setting $\mathbf{D}_k := e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}$, the update can be rewritten as affine function in $\boldsymbol{\theta}_k$,

$$\boldsymbol{\theta}_k = \Big(I - 2\alpha_k\mathbf{D}_k(\mathbf{U}_k' - \mathbf{U}_k)^\intercal\mathbf{X}_k\mathbf{X}_k^\intercal\Big)\boldsymbol{\theta}_{k-1} + \alpha_k\Big(2Y_k\mathbf{X}_k^\intercal(\mathbf{U}_k' - \mathbf{U}_k) + (\mathbf{X}_k^\intercal\mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal\mathbf{U}_k')^2\Big)\mathbf{D}_k,$$

or

$$\begin{aligned}\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star \;=\; & \Big(I - 2\alpha_k \mathbf{D}_k\big(\mathbf{U}_k' - \mathbf{U}_k\big)^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal\Big)(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^\star) \\ & + \; 2\alpha_k (Y_k - \mathbf{X}_k^\intercal \boldsymbol{\theta}^\star)\mathbf{X}_k^\intercal(\mathbf{U}_k' - \mathbf{U}_k)\mathbf{D}_k \\ & + \; \alpha_k \left((\mathbf{X}_k^\intercal \mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal \mathbf{U}_k')^2\right)\mathbf{D}_k.\end{aligned}$$

Noticing that $\varepsilon_k = Y_k - \mathbf{X}_k^\intercal \boldsymbol{\theta}^\star$, (7) follows. □

Proof of (8)*:* We show that $\mathbb{E}[\boldsymbol{\xi}_k] = 0$. By definition $\boldsymbol{\xi}_k := 2\alpha_k \varepsilon_k \mathbf{X}_k^\intercal(\mathbf{U}_k' - \mathbf{U}_k)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big) + \alpha_k \left((\mathbf{X}_k^\intercal \mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal \mathbf{U}_k')^2\right)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big)$. The first term has expectation zero, since $\mathbb{E}[\varepsilon_k] = 0$ and ε_k is independent of all the other variables. Since $\mathbf{U}_k$ and $-\mathbf{U}_k$ have the same distribution, $\mathbb{E}[e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}] = 0$ and $\mathbb{E}[(\mathbf{X}_k^\intercal \mathbf{U}_k)^2\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big)] = 0$. Using the independence of $\mathbf{U}_k, \mathbf{U}_k'$,

$$\mathbb{E}[\boldsymbol{\xi}_k] = \mathbb{E}\big[\alpha_k \left((\mathbf{X}_k^\intercal \mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal \mathbf{U}_k')^2\right)\big(e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}\big)\big] = 0.$$

A.1 Expectations with respect to the uniform distribution

The moments of the noise U play an important role in the analysis.

Definition 1. *For natural numbers $r, q \geq 0$, we abbreviate*

$$c_{r,q} \;:=\; \mathbb{E}\left[U^r (e^{-U} - e^U)^q\right] \quad \textit{with} \;\; U \sim \mathrm{Unif}[-A, A].$$

We have $c_{0,0} = 1$ and $c_{2,0} = A^2/3$. By definition, $-U(e^{-U} - e^U)$ is a non-negative random variable and therefore $-c_{1,1} > 0$. More specifically,

Lemma 3. *If $U \sim \mathrm{Unif}([-A, A])$ and $A \leq 1$, then*

$$\frac{A^2}{11} \leq -c_{1,1} \leq A^2 \quad \textit{and} \;\; c_{0,2} \leq -3c_{1,1}. \tag{21}$$

Proof. Integration by parts gives

$$\begin{aligned}-\mathbb{E}[U(e^{-U} - e^U)] &= -\frac{1}{2A}\int_{-A}^{A} u\big(e^{-u} - e^u\big)\,du = \frac{1}{A}\int_{-A}^{A} u e^u\,du = \frac{1}{A} u e^u\,\Big|_{-A}^{A} - \frac{1}{A}\int_{-A}^{A} e^u\,du \\ &= e^A + e^{-A} - \frac{e^A - e^{-A}}{A}.\end{aligned}$$

For the second part observe that for $A \geq 0$, third order Taylor expansion gives $e^A \geq 1 + A + A^2/2$ and $e^{-A} \leq 1 - A + A^2/2$. From the expression above and using that $A \leq 1$,

$$\begin{aligned}-\mathbb{E}[U(e^{-U} - e^U)] &= e^A + e^{-A} - \frac{e^A - e^{-A}}{A} \\ &= -e^A\Big(\frac{1}{A} - 1\Big) + e^{-A}\Big(1 + \frac{1}{A}\Big) \\ &\leq -\Big(1 + A + \frac{A^2}{2}\Big)\frac{1 - A}{A} + \Big(1 - A + \frac{A^2}{2}\Big)\frac{A + 1}{A} \\ &= -(1 + A)\frac{1 - A}{A} - \frac{A - A^2}{2} + (1 - A)\frac{A + 1}{A} + \frac{A^2 + A}{2} \\ &= A^2.\end{aligned}$$

By third order Taylor expansion and $A \le 1$, we find $e^A \le 1 + A + A^2/2 + e^A A^3/6 \le 1 + A + A^2/2 + eA^3/6$ and $e^{-A} \ge 1 - A + A^2/2 - eA^3/6$. Basically following the same steps as for the upper bound of $-\mathbb{E}[U(e^{-U} - e^U)]$,

$$\begin{aligned}
-\mathbb{E}[U(e^{-U} - e^U)] &= -e^A\Big(\frac{1}{A} - 1\Big) + e^{-A}\Big(1 + \frac{1}{A}\Big) \\
&\ge -\Big(1 + A + \frac{A^2}{2} + e\frac{A^3}{6}\Big)\frac{1-A}{A} + \Big(1 - A + \frac{A^2}{2} - e\frac{A^3}{6}\Big)\frac{A+1}{A} \\
&= A^2\Big(1 - \frac{e}{3}\Big).
\end{aligned}$$

Since $1 - e/3 \ge 0.093 \ge 1/11$, this completes the proof for (21).

To prove $c_{0,2} \le -3c_{1,1}$, we use that $e^x - e^{-x} = 2\sum_{\ell \text{ odd}} x^\ell/\ell!$. Since for odd ℓ, $x^\ell(e^x - e^{-x}) \ge 0$, we have for $|x| \le 1$ that $x^\ell(e^x - e^{-x}) \le x(e^x - e^{-x})$. With $e \le 3$, we obtain for $x \le 1$,

$$\begin{aligned}
\big(e^x - e^{-x}\big)^2 &= 2\sum_{\ell \text{ odd}} \frac{x^\ell}{\ell!}\big(e^x - e^{-x}\big) \le 2\sum_{\ell \text{ odd}} \frac{1}{\ell!} x\big(e^x - e^{-x}\big) \\
&\le 2\Big(e - 1 - \frac{1}{2!}\Big)x\big(e^x - e^{-x}\big) \le 3x\big(e^x - e^{-x}\big).
\end{aligned}$$

Thus $c_{0,2} = \mathbb{E}[(e^U - e^{-U})^2] \le 3\mathbb{E}[U(e^U - e^{-U})] = -3c_{1,1}$. □

Using the previous lemma,

$$c_{r,2} \le A^r c_{0,2} \le 3A^{r+2}. \tag{22}$$

For odd r, $c_{r,2} = 0$ and this inequality makes only sense if r is even.

A.2 Moments of the noise contributions in the zeroth-order scheme

Recall that $\|\cdot\|$ denotes the spectral norm for matrices and the Euclidean norm for vectors.

Lemma 4. *Let* $\mathbf{a}, \mathbf{b}$ *be column vectors of the same length. Then* $\|\mathbf{a}\mathbf{b}^\intercal\| = \|\mathbf{a}\|\|\mathbf{b}\|$.

Proof. The matrix $\mathbf{b}\mathbf{b}^\intercal$ is of rank one with non-zero eigenvalue $\mathbf{b}^\intercal\mathbf{b}$ and corresponding eigenvector $\|\mathbf{b}\|^{-1}\mathbf{b}$. Thus, $\|\mathbf{a}\mathbf{b}^\intercal\|^2 = \lambda_{\max}(\mathbf{b}\mathbf{a}^\intercal\mathbf{a}\mathbf{b}^\intercal) = \mathbf{a}^\intercal\mathbf{a}\lambda_{\max}(\mathbf{b}\mathbf{b}^\intercal) = \mathbf{a}^\intercal\mathbf{a}\mathbf{b}^\intercal\mathbf{b}$. The result follows by taking square roots. □

We now derive closed-form expressions and bounds for the two expected values in (10). Recall that $c_{r,q} = \mathbb{E}\left[U^r(e^{-U} - e^U)^q\right]$. The k-th power $\mathbf{x}^k$ of a vector $\mathbf{x} = (x_1, \ldots, x_d)$ is understood componentwise, e.g., $\|\mathbf{x}^2\|^2 = \sum_i x_i^4$.

Lemma 5. *Let* I *denote the* $d \times d$ *identity matrix. We have*

$$V(\mathbf{x}) = (c_{2,2} - c_{2,0}c_{0,2} - 2c_{1,1}^2)\operatorname{diag}(\mathbf{x}^2) + 2c_{1,1}^2\mathbf{x}\mathbf{x}^\intercal + 2c_{2,0}c_{0,2}\|\mathbf{x}\|^2 I.$$

and

$$\begin{aligned} W \;=\; \mathbb{E}\Big[&(c_{4,2} - 6c_{2,0}c_{2,2} + 6c_{2,0}^2c_{0,2} - c_{4,0}c_{0,2} - 8c_{3,1}c_{1,1} + 24c_{2,0}c_{1,1}^2)\,\mathrm{diag}(\mathbf{X}^4) \\ &+\; (4c_{2,0}c_{2,2} - 4c_{2,0}^2c_{0,2} - 8c_{2,0}c_{1,1}^2)\|\mathbf{X}\|^2\,\mathrm{diag}(\mathbf{X}^2) \\ &+\; (4c_{3,1}c_{1,1} - 12c_{2,0}c_{1,1}^2)\big(\mathbf{X}^3\mathbf{X}^\intercal + \mathbf{X}(\mathbf{X}^3)^\intercal\big) \\ &+\; 8c_{2,0}c_{1,1}^2\|\mathbf{X}\|^2\mathbf{X}\mathbf{X}^\intercal \\ &+\; (2c_{4,0}c_{0,2} - 6c_{2,0}^2c_{0,2})\|\mathbf{X}^2\|^2 I \\ &+\; 4c_{2,0}^2c_{0,2}\|\mathbf{X}\|^4 I\Big]. \end{aligned}$$

Moreover for $\mathbf{X} = (X_1, \ldots, X_d)^\intercal$,

$$\mathbb{E}[V(\mathbf{X})] \le 12A^4 d \max_{i=1,\ldots,d} \mathbb{E}[X_i^2] I, \tag{23}$$

$$\mathbb{E}\left[\|\mathbf{X}\|^2 V(\mathbf{X})\right] \le 12A^4 d^2 \max_{i=1,\ldots,d} \mathbb{E}[X_i^4] I, \tag{24}$$

$$W \le 107A^6 d^2 \max_{i=1,\ldots,d} 1 \vee \mathbb{E}[X_i^4] I. \tag{25}$$

While the closed-form expression of W depends on fourth power of $\mathbf{X}$, it is convenient to relate this to a sixth power in (25).

Proof. We first prove the formula for W. Let $\mathbf{x} = (x_1, \ldots, x_d)$ be fixed. Using that $\mathbb{E}[(\mathbf{x}^\intercal \mathbf{U}')^2] = \mathbb{E}[(\mathbf{U}')^\intercal \mathbf{x}\mathbf{x}^\intercal \mathbf{U}'] = c_{2,0}\,\mathrm{tr}(\mathbf{x}\mathbf{x}^\intercal) = c_{2,0}\mathbf{x}^\intercal\mathbf{x}$,

$$\begin{aligned} \mathbb{E}[(\mathbf{x}^\intercal \mathbf{U}')^4] &= \mathbb{E}\Big[\sum_{i,j,k,\ell} U_i' U_j' U_k' U_\ell' x_i x_j x_k x_\ell\Big] \\ &= c_{4,0}\sum_i x_i^4 + 3c_{2,0}^2 \sum_{i\neq j} x_i^2 x_j^2 \\ &= (c_{4,0} - 3c_{2,0}^2)\|\mathbf{x}^2\|^2 + 3c_{2,0}^2\|\mathbf{x}\|^4, \end{aligned}$$

$\mathbb{E}[\mathbf{D}\mathbf{D}^\intercal] = c_{0,2} I$, and the independence of $\mathbf{U}$ and $\mathbf{U}'$ gives

$$\begin{aligned} &\mathbb{E}\left[\left((\mathbf{x}^\intercal\mathbf{U})^2 - (\mathbf{x}^\intercal\mathbf{U}')^2\right)^2 \mathbf{D}\mathbf{D}^\intercal\right] \\ &=\; \mathbb{E}\left[\left((\mathbf{x}^\intercal\mathbf{U})^4 - 2c_{2,0}(\mathbf{x}^\intercal\mathbf{U})^2\mathbf{x}^\intercal\mathbf{x} + \left(c_{4,0} - 3c_{2,0}^2\right)\|\mathbf{x}^2\|^2 + 3c_{2,0}^2\|\mathbf{x}\|^4\right)\mathbf{D}\mathbf{D}^\intercal\right] \\ &=\; \mathbb{E}\left[(\mathbf{x}^\intercal\mathbf{U})^4 \mathbf{D}\mathbf{D}^\intercal\right] - 2c_{2,0}\mathbf{x}^\intercal\mathbf{x}\,\mathbb{E}\left[(\mathbf{x}^\intercal\mathbf{U})^2\mathbf{D}\mathbf{D}^\intercal\right] + c_{0,2}\left(\left(c_{4,0} - 3c_{2,0}^2\right)\|\mathbf{x}^2\|^2 + 3c_{2,0}^2\|\mathbf{x}\|^4\right) I. \end{aligned}$$

Next we simplify these two remaining expectations. The (i,j)-th off-diagonal entry of the matrix $\mathbb{E}\left[(\mathbf{x}^\intercal\mathbf{U})^2\mathbf{D}\mathbf{D}^\intercal\right]$ is $\mathbb{E}[\sum_{\ell,k} U_\ell U_k x_\ell x_k D_i D_j]$. The summands are non-zero if either $(\ell,k) = (i,j)$ or $(\ell,k) = (j,i)$. In both cases we get the contribution $c_{1,1}^2 x_i x_j$, such that for $i \neq j$, $\mathbb{E}[\sum_{\ell,k} U_\ell U_k x_\ell x_k D_i D_j] = 2c_{1,1}^2 x_i x_j$. For the i-th diagonal entry of $\mathbb{E}\left[(\mathbf{x}^\intercal\mathbf{U})^2\mathbf{D}\mathbf{D}^\intercal\right]$, we obtain $\mathbb{E}[\sum_{\ell,k} U_\ell U_k x_\ell x_k D_i^2] = \mathbb{E}[\sum_\ell U_\ell^2 x_\ell^2 D_i^2] = c_{2,2}x_i^2 + c_{2,0}c_{0,2}\sum_{\ell\neq i} x_\ell^2 = (c_{2,2} - c_{2,0}c_{0,2})x_i^2 + c_{2,0}c_{0,2}\|\mathbf{x}\|^2$. Combining these formulas yields

$$\mathbb{E}\left[(\mathbf{x}^\intercal\mathbf{U})^2\mathbf{D}\mathbf{D}^\intercal\right] = \left(c_{2,2} - c_{2,0}c_{0,2} - 2c_{1,1}^2\right)\mathrm{diag}(\mathbf{x}^2) + 2c_{1,1}^2\mathbf{x}\mathbf{x}^\intercal + c_{2,0}c_{0,2}\,\|\mathbf{x}\|_2^2\, I \tag{26}$$

and

$$\begin{aligned}\mathbb{E}\left[(\mathbf{x}^\intercal\mathbf{U})^4\mathbf{D}\mathbf{D}^\intercal\right] &= \mathbb{E}\left[\sum_{i,j,\ell,m} x_iU_ix_jU_jx_\ell U_\ell x_mU_m\mathbf{D}\mathbf{D}^\intercal\right]\\ &= (c_{4,2}-6c_{2,0}c_{2,2}+6c_{2,0}^2c_{0,2}-c_{4,0}c_{0,2}-8c_{3,1}c_{1,1}+24c_{2,0}c_{1,1}^2)\operatorname{diag}(\mathbf{x}^4)\\ &\quad+(6c_{2,0}c_{2,2}-6c_{2,0}^2c_{0,2}-12c_{2,0}c_{1,1}^2)\|\mathbf{x}\|^2\operatorname{diag}(\mathbf{x}^2)\\ &\quad+(c_{4,0}c_{0,2}-3c_{2,0}^2c_{0,2})\|\mathbf{x}^2\|^2I\\ &\quad+3c_{2,0}^2c_{0,2}\|\mathbf{x}\|^4I\\ &\quad+12c_{2,0}c_{1,1}^2\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\intercal\\ &\quad+(4c_{3,1}c_{1,1}-12c_{2,0}c_{1,1}^2)\big(\mathbf{x}^3\mathbf{x}^\intercal+\mathbf{x}(\mathbf{x}^3)^\intercal\big).\end{aligned}$$

This is because on a diagonal entry (p,p), we have $c_{4,2}x_p^4$, as well as $6c_{2,0}c_{2,2}x_p^2\sum_{q\neq p}x_q^2$ and $3c_{2,0}^2c_{0,2}(\sum_{q\neq p}x_q^2)^2$ and $c_{4,0}c_{0,2}\sum_{q\neq p}x_q^4$. On an off-diagonal entry (p,q), we have $12c_{2,0}c_{1,1}^2x_px_q\sum_{r\notin\{p,q\}}x_r^2$ as well as $4c_{3,1}c_{1,1}(x_p^3x_q+x_px_q^3)$. Combining/grouping terms, replacing $\mathbf{x}$ by the random vector $\mathbf{X}$ and taking expectation with respect to $\mathbf{X}$ yields the formula for W.

For $V(\mathbf{x})$, observe that the cross-terms are zero and by conditioning first on $\mathbf{U}'$ and using that $\mathbb{E}[\mathbf{D}\mathbf{D}^\intercal]=c_{0,2}I$, we find

$$\begin{aligned}V(\mathbf{x}) &= \mathbb{E}\left[\mathbf{x}^\intercal(\mathbf{U}'-\mathbf{U})\mathbf{D}\mathbf{D}^\intercal(\mathbf{U}'-\mathbf{U})^\intercal\mathbf{x}\right]\\ &= \mathbb{E}\left[\mathbf{x}^\intercal\mathbf{U}'\mathbf{D}\mathbf{D}^\intercal(\mathbf{U}')^\intercal\mathbf{x}\right]+\mathbb{E}\left[\mathbf{x}^\intercal\mathbf{U}\mathbf{D}\mathbf{D}^\intercal\mathbf{U}^\intercal\mathbf{x}\right]\\ &= c_{0,2}\mathbf{x}^\intercal\,\mathbb{E}\left[\mathbf{U}'(\mathbf{U}')^\intercal\right]\mathbf{x}+\mathbb{E}\left[\mathbf{x}^\intercal\mathbf{U}\mathbf{D}\mathbf{D}^\intercal\mathbf{U}^\intercal\mathbf{x}\right]\\ &= c_{0,2}c_{2,0}\|\mathbf{x}\|^2I+\mathbb{E}\left[\mathbf{x}^\intercal\mathbf{U}\mathbf{D}\mathbf{D}^\intercal\mathbf{U}^\intercal\mathbf{x}\right].\end{aligned}$$

Combined with (26), the formula $V(\mathbf{x})=(c_{2,2}-c_{2,0}c_{0,2}-2c_{1,1}^2)\operatorname{diag}(\mathbf{x}^2)+2c_{1,1}^2\mathbf{x}\mathbf{x}^\intercal+2c_{2,0}c_{0,2}\|\mathbf{x}\|^2I$ follows.

For the bounds on the expectations, we use the bounds on the moments $c_{r,q}$ derived in Section A.1. In particular $c_{r,q}\geq 0$ whenever r and q are even. The matrix $\mathbf{x}\mathbf{x}^\intercal$ is positive semi-definite. By (21), $c_{1,1}\leq A^2$ and $c_{0,2}\leq 3A^2$. Thus, $V(\mathbf{x})\leq 3A^4\operatorname{diag}(\mathbf{x}^2)+2A^4\mathbf{x}\mathbf{x}^\intercal+6A^4\|\mathbf{x}\|^2I$. Moreover, by Lemma 4, the largest eigenvalue of $\mathbf{x}\mathbf{x}^\intercal$ is $\|\mathbf{x}\|^2$ and therefore $\mathbf{x}\mathbf{x}^\intercal\leq\|\mathbf{x}\|^2I$. Since also $\operatorname{diag}(\mathbf{x}^2)\leq\|\mathbf{x}\|^2I$, we obtain $V(\mathbf{x})\leq 12A^4\|\mathbf{x}\|^2I$ and thus for $\mathbf{X}=(X_1,\dots,X_d)^\intercal$, $\mathbb{E}[V(\mathbf{X})]\leq 12A^4d\max_{i=1,\dots,d}\mathbb{E}[X_i^2]I$, proving (23). Since $X_i^2X_j^2\leq X_i^4/2+X_j^4/2$, we can derive moreover $\mathbb{E}[\|\mathbf{X}\|^2V(\mathbf{X})]\leq 12A^4d^2\max_{i,j=1,\dots,d}\mathbb{E}[X_i^2X_j^2]\leq 12A^4d^2\max_{i=1,\dots,d}\mathbb{E}[X_i^4]$, proving (24).

We finally derive (25). Since $2\|\mathbf{x}^3\mathbf{x}^\intercal\|=2\|\mathbf{x}^3\|\|\mathbf{x}\|$, all eigenvalues of $\mathbf{x}^3\mathbf{x}^\intercal+\mathbf{x}(\mathbf{x}^3)^\intercal$ lie between $-2\|\mathbf{x}^3\|\|\mathbf{x}\|$ and $2\|\mathbf{x}^3\|\|\mathbf{x}\|$. Therefore, for any real number a, we have $a(\mathbf{x}^3\mathbf{x}^\intercal+\mathbf{x}(\mathbf{x}^3)^\intercal)\leq 2|a|\|\mathbf{x}^3\|\|\mathbf{x}\|$. By (22), $c_{r,2}\leq A^rc_{0,2}\leq 3A^{r+2}$. Moreover $0\leq c_{3,1}c_{1,1}\leq A^2c_{1,1}^2\leq A^6$ and $0\leq c_{2,0}c_{1,1}^2\leq A^6$. Thus,

$$\begin{aligned}W\leq&\mathbb{E}\Big[(3+18+24)A^6\operatorname{diag}(\mathbf{X}^4)+12A^6\|\mathbf{X}\|^2\operatorname{diag}(\mathbf{X}^2)+12A^62\|\mathbf{X}^3\|\|\mathbf{X}\|I+8A^6\|\mathbf{X}\|^4I\\ &+6A^6\|\mathbf{X}^2\|^2I+12A^6\|\mathbf{X}\|^4I\Big].\end{aligned}$$

Since $\mathrm{diag}(\mathbf{X}^4) \le \|\mathbf{X}\|^4 I$, $\mathrm{diag}(\mathbf{X}^2) \le \|\mathbf{X}\|^2 I$, and $\|\mathbf{X}^2\|^2 \le \|\mathbf{X}\|^4$, we obtain $W \le 83A^6\mathbb{E}[\|\mathbf{X}\|^4]I + 24A^6\mathbb{E}[\|\mathbf{X}^3\|\|\mathbf{X}\|]I$.

Because of the elementary inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$,

$$\mathbb{E}\left[\|\mathbf{X}\|^4\right] = \sum_{i,j=1}^{d} \mathbb{E}[X_i^2 X_j^2] \le \frac{1}{2}\sum_{i,j=1}^{d} \mathbb{E}[X_i^4] + \mathbb{E}[X_j^4] \le d^2 \max_{i=1,\dots,d} \mathbb{E}[X_i^4].$$

Similarly, $\|\mathbf{X}^3\|\|\mathbf{X}\| \le \sqrt{d\max_i X_i^6}\sqrt{d\max_i X_i^2} = d\max_i X_i^4 \le d\sum_i X_i^4$ and hence

$$\mathbb{E}\left[\|\mathbf{X}^3\|\|\mathbf{X}\|\right] \le d\sum_i \mathbb{E}[X_i^4] \le d^2 \max_i \mathbb{E}[X_i^4].$$

Combined with $W \le 83A^6\mathbb{E}[\|\mathbf{X}\|^4]I + 24A^6\mathbb{E}[\|\mathbf{X}^3\|\|\mathbf{X}\|]I$, we finally obtain

$$W \le 107A^6 d^2 \max_{i=1,\dots,d} 1 \vee \mathbb{E}[X_i^4] I.$$

□

A.3 Proof of Lemma 1

Recall that $\mathbf{D}_k := e^{-\mathbf{U}_k} - e^{\mathbf{U}_k}$. The rewritten update equation (7) can be decomposed into the following three parts

$$\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star = \mathbf{A} + \mathbf{B} + \mathbf{C}, \tag{27}$$

where

$$\begin{aligned}
\mathbf{A} &:= (I - 2\alpha_k \mathbf{D}_k(\mathbf{U}_k' - \mathbf{U}_k)^\intercal \mathbf{X}_k \mathbf{X}_k^\intercal)(\boldsymbol{\theta}_{k-1} - \boldsymbol{\theta}^\star),\\
\mathbf{B} &:= 2\alpha_k \varepsilon_k \mathbf{X}_k^\intercal(\mathbf{U}_k' - \mathbf{U}_k)\mathbf{D}_k,\\
\mathbf{C} &:= \alpha_k\left((\mathbf{X}_k^\intercal \mathbf{U}_k)^2 - (\mathbf{X}_k^\intercal \mathbf{U}_k')^2\right)\mathbf{D}_k.
\end{aligned}$$

In the following we often use that $(\mathbf{X}_k, Y_k)$ has the same distribution as $(\mathbf{X}, Y)$. This means that if all randomness in an expectation is only due to $(\mathbf{X}_k, Y_k)$, these variables can be replaced by $(\mathbf{X}, Y)$. Recall moreover that odd powers of $\mathbf{U}$ or $\mathbf{U}'$ disappear and that we have defined $\mu = -c_{1,1}$. Thus,

$$\begin{aligned}
\mathbb{E}[\mathbf{A}\mathbf{A}^\intercal] &= S_{k-1} - 2\mu\alpha_k(QS_{k-1} + S_{k-1}Q) + 4\alpha_k^2 \mathbb{E}\left[\mathbf{X}_k^\intercal S_{k-1}\mathbf{X}_k V(\mathbf{X}_k)\right]\\
&= (I - 2\alpha_k\mu Q) S_{k-1} (I - 2\alpha_k \mu Q) + 4\alpha_k^2\left\{\mathbb{E}\left[\mathbf{X}^\intercal S_{k-1}\mathbf{X} V(\mathbf{X})\right] - \mu^2 Q S_{k-1} Q\right\}.
\end{aligned} \tag{28}$$

Since for any $d \times d$ matrix T, we have $T^\intercal S_{k-1} T \le \|S_{k-1}\| T^\intercal T$, we also have

$$\mathbb{E}[\mathbf{A}\mathbf{A}^\intercal] \le \|S_{k-1}\|\left(I - 4\alpha_k\mu Q + 4\alpha_k^2 \mathbb{E}\left[\|\mathbf{X}\|^2 V(\mathbf{X})\right]\right). \tag{29}$$

Now by counting powers of $\mathbf{U}_k, \mathbf{U}_k', \mathbf{D}_k$, we find that

$$\mathbb{E}[\mathbf{A}\mathbf{C}^\intercal] = 0.$$

Applying tower rule by conditioning on all randomness except $\varepsilon_k \sim \mathcal{N}(0, \sigma^2)$, it follows that

$$\begin{aligned}
\mathbb{E}[\mathbf{A}\mathbf{B}^\intercal] &= \mathbb{E}\left[\mathbb{E}[\mathbf{A}\mathbf{B}^\intercal \,|\, \mathbf{U}_k, \mathbf{U}_k', \mathbf{X}_k, \boldsymbol{\theta}_k]\right] = \mathbb{E}\left[\mathbf{A} 2\alpha_k \mathbf{D}_k^\intercal (\mathbf{U}_k' - \mathbf{U}_k)^\intercal \mathbf{X}_k \,\mathbb{E}[\varepsilon_k]\right] = 0, \\
\mathbb{E}[\mathbf{B}\mathbf{C}^\intercal] &= 0, \\
\mathbb{E}[\mathbf{B}\mathbf{B}^\intercal] &= 4\alpha_k^2 \,\mathbb{E}\left[\varepsilon_k^2 V(\mathbf{X})\right] = 4\alpha_k^2 \sigma^2 \,\mathbb{E}\left[V(\mathbf{X})\right],
\end{aligned}$$

and finally

$$\mathbb{E}[\mathbf{C}\mathbf{C}^\intercal] = \alpha_k^2 W,$$

which is some combination of fourth powers of $\mathbf{X}$.

Now (i) follows due to $\mathbb{E}[G_k \mathbf{W}_k \boldsymbol{\xi}_k^\intercal] = \mathbb{E}[\mathbf{A}(\mathbf{B} + \mathbf{C})^\intercal] = 0$. Statement (ii) follows since by (8), $\mathbb{E}[\boldsymbol{\xi}_k] = 0$ such that $\mathrm{Cov}(\boldsymbol{\xi}_k) = \mathbb{E}[\boldsymbol{\xi}_k \boldsymbol{\xi}_k^\intercal] = \mathbb{E}[(\mathbf{B} + \mathbf{C})(\mathbf{B} + \mathbf{C})^\intercal] = \mathbb{E}[\mathbf{B}\mathbf{B}^\intercal] + \mathbb{E}[\mathbf{C}\mathbf{C}^\intercal] = 4\alpha_k^2 \sigma^2 \,\mathbb{E}\left[V(\mathbf{X})\right] + \alpha_k^2 W$. To see (iii), observe that

$$\begin{aligned}
S_k &= \mathbb{E}[(\mathbf{A} + \mathbf{B} + \mathbf{C})(\mathbf{A} + \mathbf{B} + \mathbf{C})^\intercal] \\
&= \mathbb{E}\left[\mathbf{A}\mathbf{A}^\intercal\right] + \mathbb{E}\left[\mathbf{A}\mathbf{B}^\intercal\right] + \mathbb{E}\left[\mathbf{B}\mathbf{A}^\intercal\right] + \mathbb{E}\left[\mathbf{B}\mathbf{B}^\intercal\right] + \mathbb{E}\left[\mathbf{C}\mathbf{C}^\intercal\right] \\
&= \big(I - 2\alpha_k \mu Q\big) S_{k-1} \big(I - 2\alpha_k \mu Q\big) + 4\alpha_k^2 \Big(\mathbb{E}\left[\mathbf{X}^\intercal S_{k-1} \mathbf{X} V(\mathbf{X})\right] - \mu^2 Q S_{k-1} Q \Big) \\
&\quad + 4\alpha_k^2 \sigma^2 \mathbb{E}[V(\mathbf{X})] + \alpha_k^2 W.
\end{aligned}$$

Applying (29) instead of (28) yields the asserted inequality. □

A.4 Proof of Theorem 1

We now combine the recursive formula with the bounds obtained in (23), (24), and (25). Recall that by definition $\mu = -c_{1,1}$ and by Lemma 3, $\mu \geq A^2/11$ which is the same as $-\mu \leq -A^2/11$. Setting $M_k := \max_{i=1,\ldots,d} 1 \vee E[X_i^k]$, we obtain

$$\|S_k\| \leq \|S_{k-1}\| \Big(1 - \frac{4}{11} \alpha_k A^2 \lambda_{\min}(Q) + 48 \alpha_k^2 A^4 d^2 M_4\Big) + 48 \alpha_k^2 \sigma^2 A^4 d M_2 + \alpha_k^2 107 A^6 d^2 M_4.$$

By assumption $A = \sigma/\sqrt{d}$. If

$$96 \alpha_k^2 A^4 d^2 M_4 \leq \frac{4}{11} \alpha_k A^2 \lambda_{\min}(Q) \leq 1, \tag{30}$$

then, the previous inequality yields the simpler recursion inequality

$$\begin{aligned}
\|S_k\| &\leq \|S_{k-1}\| \Big(1 - \frac{2}{11} \alpha_k A^2 \lambda_{\min}(Q)\Big) + (48 M_2 + 107 M_4) \alpha_k^2 \sigma^2 A^4 d \\
&= \|S_{k-1}\| \big(1 - \beta_k\big) + C_* \beta_k^2 \sigma^2 d,
\end{aligned}$$

where we defined

$$\beta_k := \frac{2}{11} \alpha_k A^2 \lambda_{\min}(Q) \quad \text{and} \quad C_* := (48 M_2 + 107 M_4) \Big(\frac{11}{2 \lambda_{\min}(Q)}\Big)^2. \tag{31}$$

Induction with respect to k gives

$$\|S_k\| \le \|S_0\| \prod_{\ell=1}^{k} \big(1-\beta_\ell\big) + C_* \sum_{j=1}^{k} \Big(\prod_{\ell=j+1}^{k} \big(1-\beta_\ell\big) \Big) \beta_j^2 \sigma^2 d, \tag{32}$$

where $\prod_{\ell=k+1}^{k}(1-\beta_\ell) := 1$. Rewritten in terms of β_k, the assumed inequality (30) is

$$96\Big(\frac{11\beta_k^2}{2\lambda_{\min}(Q)}\Big)^2 d^2 M_4 \le 2\beta_k \le 1.$$

The choice of the learning rate for α_k in (12) results in

$$\beta_k = \frac{2}{11}\alpha_k A^2 \lambda_{\min}(Q) = \frac{2B\log(d)}{Bk + d^2\log(d)} \quad \text{with} \quad B = 1 \wedge \frac{\lambda_{\min}(Q)^2}{2904 M_4}. \tag{33}$$

For all $d \ge 1$, we have $\beta_k \le 2B/d^2$. Using that $96(11/2)^2 = 2904$, one can now check that for all $d \ge 2$, the assumed inequalities for β_k hold and afortiori thus also (30).

Let k^* be the smallest integer such that $Bk^* \ge d^2\log(d)$. Since $d \ge 2$, we must have

$$k^* \le \frac{1 + d^2\log(d)}{B} \le \frac{2d^2\log(d)}{B}. \tag{34}$$

For all $k \ge k^*$, we have $\beta_k \ge \log(d)/k$. By bounding $1-\beta_\ell \le 1$ for all $\ell < k^*$, and using for $k \ge j+1 \ge k^*$, $\sum_{\ell=j+1}^{k} \frac{1}{\ell} \ge \sum_{\ell=j+1}^{k} \int_{\ell}^{\ell+1} \frac{1}{\ell}\,du = \int_{j+1}^{k+1} \frac{1}{u}\,du = \log(k+1) - \log(j+1) = \log((k+1)/(j+1))$, we obtain for all $k \ge k^*$,

$$\prod_{\ell=j+1}^{k} \big(1-\beta_\ell\big) \le \exp\Big(-\log(d) \sum_{\ell=k^*\vee(j+1)}^{k} \frac{1}{\ell} \Big) \le \Big(\frac{k^* \vee (j+1)}{k+1} \Big)^{\log(d)}.$$

Combined with $\beta_k \le 2B/d^2$ for all k, $\beta_j \le \log(d)/j \le 2\log(d)/(j+1)$ for all $j \ge k^*$, and (32), we obtain for all $k > k^*$,

$$\begin{aligned}
\|S_k\| &\le \Big(\frac{k^*}{k}\Big)^{\log(d)} \|S_0\| + C_* \sum_{j=1}^{k^*-1} \Big(\frac{k^*}{k}\Big)^{\log(d)} \Big(\frac{2B}{d^2}\Big)^2 \sigma^2 d + C_* \sum_{j=k^*}^{k} \Big(\frac{j+1}{k+1}\Big)^{\log(d)} \Big(\frac{2\log(d)}{j+1}\Big)^2 \sigma^2 d \\
&\le \Big(\frac{k^*}{k}\Big)^{\log(d)} \|S_0\| + C_* \frac{1}{k} \Big(\frac{k^*}{k}\Big)^{\log(d)-1} (k^*)^2 \Big(\frac{2B}{d^2}\Big)^2 \sigma^2 d + C_* \sum_{j=k^*}^{k} \Big(\frac{j+1}{k+1}\Big)^{\log(d)} \Big(\frac{2\log(d)}{j+1}\Big)^2 \sigma^2 d.
\end{aligned}$$

Since $d \ge 9$, we have $\log(d) - 2 > 0$ and therefore $k^*/k \le 1$. Using also (34) gives

$$\frac{1}{k}\Big(\frac{k^*}{k}\Big)^{\log(d)-1} (k^*)^2 \Big(\frac{2B}{d^2}\Big)^2 \sigma^2 d \le \frac{1}{k}\Big(\frac{2d^2\log(d)}{B}\Big)^2 \Big(\frac{2B}{d^2}\Big)^2 \sigma^2 d \le \frac{16\sigma^2 d\log^2(d)}{k}.$$

Applying $\log(d) - 2 > 0$ again and moreover, $(j+1)/(k+1) \le 1$ for all $j \le k$, we find

$$\sum_{j=k^*}^{k} \Big(\frac{j+1}{k+1}\Big)^{\log(d)} \frac{1}{(j+1)^2} = \frac{1}{(k+1)^2} \sum_{j=k^*}^{k} \Big(\frac{j+1}{k+1}\Big)^{\log(d)-2} \le \frac{1}{k}.$$

Combined, the last three displayed inequalities combined yield

$$\|S_k\| \le \Big(\frac{k^*}{k}\Big)^{\log(d)} \|S_0\| + C_* \frac{20\sigma^2 d\log^2(d)}{k}.$$

The first claim follows now with $C = 20C_*$.

For two positive semi-definite matrices $D = D_1^\intercal D_1$ and E, we know that $D_1 E D_1^\intercal$ is positive semi-definite and bounded by $\|E\| D_1 D_1^\intercal$. Therefore, $\operatorname{tr}(DE) = \operatorname{tr}(D_1 E D_1^\intercal) \le \operatorname{tr}(\|E\| D_1 D_1^\intercal) = \|E\| \operatorname{tr}(D_1^\intercal D_1) = \|E\| \operatorname{tr}(D)$. Applying this to $(D, E) = (Q, S_k)$ and using that

$$\mathbb{E}\left[(\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)^\intercal Q (\boldsymbol{\theta}_k - \boldsymbol{\theta}^\star)\right] = \operatorname{tr}(QS_k)$$

as well as $\operatorname{tr}(Q) \le d$ yields the second claimed inequality. □

Proof of Inequality (14). Set $k_* := 2e^{\gamma+3\kappa} d^2 \log(d)/B$. Then,

$$\Big(\frac{2d^2\log(d)}{Bk_*}\Big)^{\log(d)} \le d^{-\gamma-3\kappa} \le \frac{1}{d^\gamma (d^2\log(d))^\kappa} \le \frac{C'}{d^\gamma k_*^\kappa}, \tag{35}$$

with $C' = (2e^{\gamma+3\kappa}/B)^\kappa$. If for positive A, B, u, v, x with $u > v$, we have $A/x^u \le B/x^v$, then also $A/y^u \le B/y^v$ for all $y \ge x$. Since $d > e^\kappa$ we can apply this inequality with $u = \log(d) > \kappa = v$, proving that (35) holds for all $k \ge k_*$.

B Proofs for the lower bounds

Proof of Lemma 2. To distinguish the two models, we rename the queries in (16) into $W_1, W_1', \ldots, W_k, W_k'$ and the query vectors into $\mathbf{w}_1, \mathbf{w}_1', \ldots, \mathbf{w}_k, \mathbf{w}_k'$. This means that the second query model is then denoted by $\{W_1, W_1', \ldots, W_k, W_k', \mathbf{w}_1, \mathbf{w}_1', \ldots, \mathbf{w}_k, \mathbf{w}_k'\}$ with $W_\ell = Y_\ell - \mathbf{X}_\ell^\intercal \mathbf{w}_\ell$ and $W_\ell' = \mathbf{X}_\ell^\intercal \mathbf{w}_\ell'$.

To see the equivalence, we apply induction with respect to k. The base case $k = 1$ and the induction step $k \to k+1$ are similar and therefore only the latter will be discussed.

The induction step $k \to k+1$ is split in two parts. We first prove that the first query model can be transformed into the second query model without knowledge of the parameters. To see this, choose query vectors $\mathbf{v}_{k+1} = \mathbf{w}_{k+1} + \mathbf{w}_{k+1}'$ and $\mathbf{v}_{k+1}' = \mathbf{w}_{k+1} - \mathbf{w}_{k+1}'$. Query vectors can depend on previously seen queries and query vectors, such that $\mathbf{w}_{k+1}, \mathbf{w}_{k+1}'$ can depend on $\{W_1, W_1', \ldots, W_k, W_k', \mathbf{w}_1, \mathbf{w}_1', \ldots, \mathbf{w}_k, \mathbf{w}_k'\}$ and $\mathbf{v}_{k+1}, \mathbf{v}_{k+1}'$ can depend on $\{Z_1, Z_1', \ldots, Z_k, Z_k', \mathbf{v}_1, \mathbf{v}_1', \ldots, \mathbf{v}_k, \mathbf{v}_k'\}$. Since the models can be transformed into each other by the induction hypothesis, those are eligible query vectors. The corresponding queries in (15) are then given by $Z_{k+1} = Y_{k+1} - \mathbf{X}_{k+1}^\intercal(\mathbf{w}_{k+1} + \mathbf{w}_{k+1}')$ and $Z_{k+1}' = Y_{k+1} - \mathbf{X}_{k+1}^\intercal(\mathbf{w}_{k+1} - \mathbf{w}_{k+1}')$. Now, $W_{k+1} := \frac{1}{2}(Z_{k+1} + Z_{k+1}') = Y_{k+1} - \mathbf{X}_{k+1}^\intercal \mathbf{w}_{k+1}$ and $W_{k+1}' := \frac{1}{2}(Z_{k+1}' - Z_{k+1}) = \mathbf{X}_{k+1}^\intercal \mathbf{w}_{k+1}'$ are then the queries from the second model. Since one can also retrieve the query vectors $\mathbf{w}_{k+1} = \frac{1}{2}(\mathbf{v}_{k+1} + \mathbf{v}_{k+1}')$ and $\mathbf{w}_{k+1}' = \frac{1}{2}(\mathbf{v}_{k+1} - \mathbf{v}_{k+1}')$, we can transform the data from the first query model into data from the second query model $\{W_1, W_1', \ldots, W_{k+1}, W_{k+1}', \mathbf{w}_1, \mathbf{w}_1', \ldots, \mathbf{w}_{k+1}, \mathbf{w}_{k+1}'\}$, completing the first part of the induction step.

The induction step for the other direction is similar and omitted. □

As in the previous work on lower bounds for sequential designs [27, 2, 12], we use a version of Assouad's lemma [30]. Recall that θ_j denotes the j-th component of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\intercal$.

Lemma 6. *For any estimator $\widehat{\boldsymbol{\theta}}$ and any $\rho > 0$,*

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\ \big[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\big] \geq \frac{\rho^2 d}{2}\Big(1-\frac{1}{d}\sum_{j=1}^d \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k},\mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big)$$

with

$$\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k} := \frac{1}{2^{d-1}} \sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d \text{ and } \theta_j=\theta_j^\star+\rho} P_{\boldsymbol{\theta},\mathcal{V}_k},$$

and

$$\mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k} := \frac{1}{2^{d-1}} \sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d \text{ and } \theta_j=\theta_j^\star-\rho} P_{\boldsymbol{\theta},\mathcal{V}_k}.$$

Proof. Let ξ be either $+1$ or -1. If $\boldsymbol{\theta}$ is an element of $\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d$ with $\theta_j=\theta_j^\star+\xi\rho$, then, $(\widehat{\theta}_j-\theta_j)^2 \geq \rho^2\mathbf{1}(\mathrm{sign}(\widehat{\theta}_j-\theta_j^\star)\neq\xi)$, where the sign function evaluated at zero is defined as $+1$. For probability measures P,Q and a measurable set A, we have $P(A)+Q(A^c)=1+Q(A^c)-P(A^c)\geq 1-\mathrm{TV}(P,Q)$. Rewriting $\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2=\sum_{j=1}^d(\widehat{\theta}_j-\theta_j)^2$, and applying the lower bound

$$\max_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d} \geq \frac{1}{2^d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d} = \frac{1}{2^d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d \text{ and } \theta_j=\theta_j^\star+\rho} + \frac{1}{2^d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d \text{ and } \theta_j=\theta_j^\star-\rho},$$

we find

$$\begin{aligned}
&\max_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\ \big[\|\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}\|^2\big]\\
&= \max_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d} \sum_{j=1}^d \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\ \big[(\widehat{\theta}_j-\theta_j)^2\big]\\
&\geq \rho^2\sum_{j=1}^d \frac{1}{2}\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\big(\mathrm{sign}(\widehat{\theta}_j-\theta_j^\star)\neq 1\big)+\frac{1}{2}\mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\big(\mathrm{sign}(\widehat{\theta}_j-\theta_j^\star)\neq -1\big)\\
&\geq \frac{\rho^2 d}{2}\Big(1-\frac{1}{d}\sum_{j=1}^d \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k},\mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big).
\end{aligned} \tag{36}$$

The right hand side does not depend anymore on the estimator $\widehat{\boldsymbol{\theta}}$. □

Proof of Theorem 2. For $\sigma^2=0$, there is nothing to prove and therefore, we assume $\sigma^2>0$.

We work in the equivalent model (16). One can assume that $\mathbf{v}'_\ell\neq 0$ for all $\ell=1,\dots,k$. Otherwise, $Z'_\ell=0$ and changing $\mathbf{v}'_\ell$ to a non-zero vector yields additional information about the model. We can also assume that for any $\ell=1,\dots,k$,

$$\|\mathbf{v}'_\ell\|=1, \tag{37}$$

as any other scaling would simply scale the observation Z'_ℓ. Define

$$\tau := \frac{R}{9\sqrt{d}}\Big(1\wedge\frac{d\sigma}{\sqrt{k}R}\Big). \tag{38}$$

One can think of τ^2 as the convergence rate of an individual component $\mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}(\widehat{\boldsymbol{\theta}}_j - \boldsymbol{\theta}_j)^2$.

We want to apply Assouad's lower bound in Lemma 6. For any $\boldsymbol{\theta}^\star \in B_{R/2}(0)$ and any $\boldsymbol{\theta} \in \boldsymbol{\theta}^\star + \{-\tau,\tau\}^d$, we have $\|\boldsymbol{\theta}\| \le \|\boldsymbol{\theta}^\star\| + R/9 \le R$. Thus, we can lower bound $\sup_{\boldsymbol{\theta}\in B_R(0)} \ge \sup_{\boldsymbol{\theta}^\star\in B_{R/2}(0)} \max_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d}$. Lemma 6 with ρ replaced by τ gives therefore

$$\sup_{\boldsymbol{\theta}\in B_R(0)} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\right] \ge \sup_{\boldsymbol{\theta}^\star\in B_{R/2}(0)} \frac{\tau^2 d}{2}\Big(1 - \frac{1}{d}\sum_{j=1}^d \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big) \tag{39}$$

with

$$\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k} := \frac{1}{2^{d-1}} \sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d \text{ and } \theta_j=\theta_j^\star+\tau} P_{\boldsymbol{\theta},\mathcal{V}_k}, \tag{40}$$

and

$$\mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k} := \frac{1}{2^{d-1}} \sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d \text{ and } \theta_j=\theta_j^\star-\tau} P_{\boldsymbol{\theta},\mathcal{V}_k}. \tag{41}$$

Pinsker's inequality $\mathrm{TV}(P,Q) \le \sqrt{\mathrm{KL}(P,Q)/2}$ combined with Cauchy-Schwarz inequality $(\frac{1}{d}\sum_{j=1}^d b_j)^2 \le (\sum_{j=1}^d d^{-2})(\sum_{j=1}^d b_j^2) = \frac{1}{d}\sum_{j=1}^d b_j^2$ and the joint convexity of the Kullback-Leibler divergence $\mathrm{KL}(\lambda_1 P_1 + \ldots + \lambda_m P_m, \lambda_1 Q_1 + \ldots + \lambda_m Q_m) \le \sum_{s=1}^m \lambda_s \mathrm{KL}(P_s, Q_s)$ for $\lambda_1,\ldots,\lambda_m \ge 0$ and $\sum_{s=1}^m \lambda_s = 1$, yield

$$\begin{aligned}
&\Big(\frac{1}{d}\sum_{j=1}^d \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big)^2 \\
&\le \frac{1}{d}\sum_{j=1}^d \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)^2 \\
&\le \frac{1}{2d}\sum_{j=1}^d \mathrm{KL}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right) \\
&\le \frac{1}{2d}\sum_{j=1}^d \frac{1}{2^{d-1}} \sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d \text{ and } \theta_j=\theta_j^\star+\tau} \mathrm{KL}\left(P_{\boldsymbol{\theta},\mathcal{V}_k}, P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,\mathcal{V}_k}\right) \\
&= \frac{1}{2^d d} \sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d} \sum_{j=1}^d \mathrm{KL}\left(P_{\boldsymbol{\theta},\mathcal{V}_k}, P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,\mathcal{V}_k}\right)\mathbf{1}\big(\theta_j = \theta_j^\star + \tau\big).
\end{aligned} \tag{42}$$

with $\mathbf{e}_j$ the j-th standard basis vector.

In a next step, we need to bound $\mathrm{KL}(P_{\boldsymbol{\theta},\mathcal{V}_k}, P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,\mathcal{V}_k})$. The chain rule for the Kullback-Leibler divergence states that

$$\begin{aligned}
\mathrm{KL}(P_{U,V}, Q_{U,V}) &= \mathbb{E}_{P_V}\left[\mathrm{KL}(P_{U|V}, Q_{U|V})\right] + \mathrm{KL}(P_V, Q_V) \\
&= \mathbb{E}_{P_{U,V}}\left[\mathrm{KL}(P_{U|V}, Q_{U|V}) + \mathrm{KL}(P_V, Q_V)\right].
\end{aligned}$$

The data are generated sequentially,

$$(\mathbf{v}_1, \mathbf{v}_1') \to (Z_1, Z_1') \to \ldots \to (\mathbf{v}_k, \mathbf{v}_k') \to (Z_k, Z_k').$$

After the query vectors $(\mathbf{v}_\ell, \mathbf{v}'_\ell)$ are selected, the queries are given by $Z_\ell = Y_\ell - \mathbf{X}_\ell^\intercal \mathbf{v}_\ell$ and $Z'_\ell = \mathbf{X}_\ell^\intercal \mathbf{v}'_\ell$. Therefore,

$$(Z_\ell, Z'_\ell)\big|(Z_1, Z'_1, \ldots, Z_{\ell-1}, Z'_{\ell-1}, \mathbf{v}_1, \mathbf{v}'_1, \ldots, \mathbf{v}_\ell, \mathbf{v}'_\ell) = (Z_\ell, Z'_\ell)\big|(\mathbf{v}_\ell, \mathbf{v}'_\ell).$$

The distribution of $(\mathbf{v}_\ell, \mathbf{v}'_\ell)|(Z_1, Z'_1, \ldots, Z_{\ell-1}, Z'_{\ell-1}, \mathbf{v}_1, \mathbf{v}'_1, \ldots, \mathbf{v}_{\ell-1}, \mathbf{v}'_{\ell-1})$ does not depend on the unknown regression vector $\boldsymbol{\theta}$. Write Q_ℓ to denote this distribution. If $P_{\boldsymbol{\theta},(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}$ denotes the distribution of $(Z_\ell, Z'_\ell)\big|(\mathbf{v}_\ell, \mathbf{v}'_\ell)$ for the data generating parameter $\boldsymbol{\theta}$, the chain rule and the arguments above yield

$$\begin{aligned}
\mathrm{KL}\left(P_{\boldsymbol{\theta},\mathcal{V}_k}, P_{\boldsymbol{\theta}',\mathcal{V}_k}\right) &= \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\Big[\sum_{\ell=1}^{k} \mathrm{KL}\left(P_{\boldsymbol{\theta},(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}, P_{\boldsymbol{\theta}',(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}\right) + \mathrm{KL}(Q_\ell, Q_\ell)\Big] \\
&= \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\Big[\sum_{\ell=1}^{k} \mathrm{KL}\left(P_{\boldsymbol{\theta},(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}, P_{\boldsymbol{\theta}',(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}\right)\Big].
\end{aligned} \tag{43}$$

We will now apply this identity for $\boldsymbol{\theta}' = \boldsymbol{\theta} - 2\tau\mathbf{e}_j$ to eventually derive a bound for (42).

For two centered d-variate normal distributions $\mathcal{N}(0, \Sigma_0)$, $\mathcal{N}(0, \Sigma_1)$, with $\Sigma_0, \Sigma_1 > 0$, we have

$$\mathrm{KL}\left(\mathcal{N}(0, \Sigma_1^2), \mathcal{N}(0, \Sigma_0^2)\right) = \frac{1}{2}\Big(\log\Big(\frac{\det(\Sigma_0)}{\det(\Sigma_1)}\Big) + \mathrm{tr}\left(\Sigma_1\Sigma_0^{-1}\right) - d\Big).$$

Consider an invertible matrix T, and let $S_0 := T\Sigma_0 T^\intercal$, $S_1 := T\Sigma_1 T^\intercal$. Assume further that there are positive definite matrices Λ_0, Λ_1 such that $S_0 \geq \Lambda_0$ and $S_1 \geq \Lambda_1$. It is known (e.g. Theorem 6.8 in [31]), that $S_0^{-1} \leq \Lambda_0^{-1}$, $S_1^{-1} \leq \Lambda_1^{-1}$, and that $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ for square matrices A, B. Using these facts, the symmetrized Kullback-Leibler divergence can be bounded as follows

$$\begin{aligned}
\mathrm{KL}\left(\mathcal{N}(0, \Sigma_0^2), \mathcal{N}(0, \Sigma_1^2)\right) &\leq \mathrm{KL}\left(\mathcal{N}(0, \Sigma_0^2), \mathcal{N}(0, \Sigma_1^2)\right) + \mathrm{KL}\left(\mathcal{N}(0, \Sigma_1^2), \mathcal{N}(0, \Sigma_0^2)\right) \\
&= \frac{1}{2}\Big(\mathrm{tr}\left(\Sigma_1\Sigma_0^{-1}\right) + \mathrm{tr}\left(\Sigma_0\Sigma_1^{-1}\right) - 2d\Big) \\
&= \frac{1}{2}\,\mathrm{tr}\left(\Sigma_1^{-1}(\Sigma_1 - \Sigma_0)\Sigma_0^{-1}(\Sigma_1 - \Sigma_0)\right) \\
&= \frac{1}{2}\,\mathrm{tr}\left((T\Sigma_1 T^\intercal)^{-1}(T\Sigma_1 T^\intercal - T\Sigma_0 T^\intercal)(T\Sigma_0 T^\intercal)^{-1}(T\Sigma_1 T^\intercal - T\Sigma_0 T^\intercal)\right) \\
&= \frac{1}{2}\,\mathrm{tr}\left(S_1^{-1}(S_1 - S_0)S_0^{-1}(S_1 - S_0)\right) \\
&= \frac{1}{2}\,\mathrm{tr}\left(S_1^{-1/2}(S_1 - S_0)S_0^{-1}(S_1 - S_0)S_1^{-1/2}\right) \\
&\leq \frac{1}{2}\,\mathrm{tr}\left(S_1^{-1/2}(S_1 - S_0)\Lambda_0^{-1}(S_1 - S_0)S_1^{-1/2}\right) \\
&= \frac{1}{2}\,\mathrm{tr}\left(\Lambda_0^{-1/2}(S_1 - S_0)S_1^{-1}(S_1 - S_0)\Lambda_0^{-1/2}\right) \\
&\leq \frac{1}{2}\,\mathrm{tr}\left(\Lambda_0^{-1/2}(S_1 - S_0)\Lambda_1^{-1}(S_1 - S_0)\Lambda_0^{-1/2}\right) \\
&= \frac{1}{2}\,\mathrm{tr}\left(\Lambda_0^{-1}(S_1 - S_0)\Lambda_1^{-1}(S_1 - S_0)\right).
\end{aligned} \tag{44}$$

We now control the right hand side of (43) for $\boldsymbol{\theta}' = \boldsymbol{\theta} - 2\tau\mathbf{e}_j$. As we work in the equivalent model (16) with the normalization constraint (37), the distributions of $(Z_\ell, Z'_\ell)^\intercal|(\mathbf{v}_\ell, \mathbf{v}'_\ell)$ is given by (17).

We have normalized $\mathbf{v}'_\ell$ to a unit-length vector $\|\mathbf{v}'_\ell\| = 1$. Thus, the distributions $P_{\boldsymbol{\theta},(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}$ and $P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}$ (as shown in (17)) are centered normal with respective covariances

$$\begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell\|^2 & \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \\ \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell\|^2 & \langle \boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \\ \langle \boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle & 1 \end{pmatrix}.$$

In a next step, we transform the model such that the subsequent analysis becomes more tractable. Choosing the transformation

$$T = \begin{pmatrix} 1 & -\langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \\ 0 & 1 \end{pmatrix}$$

maps $(Z_\ell, Z'_\ell)^\intercal$ to $(Z_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle Z'_\ell, Z'_\ell)^\intercal$ transforming the covariances via the identities

$$S_0 := T \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell\|^2 & \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \\ \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle & 1 \end{pmatrix} T^\intercal = \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2 & 0 \\ 0 & 1 \end{pmatrix} \tag{45}$$

and

$$\begin{aligned} S_1 &:= T \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell\|^2 & \langle \boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \\ \langle \boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle & 1 \end{pmatrix} T^\intercal \\ &= \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2 & -2\tau\langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle \\ -2\tau\langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle & 1 \end{pmatrix}. \end{aligned} \tag{46}$$

For a symmetric 2×2 matrix, the elementary inequality $2ab \geq -|2ab| \geq -2a^2 - \frac{1}{2}b^2$ yields for all vectors $(u_1, u_2)^\intercal$,

$$(u_1 u_2)^\intercal \begin{pmatrix} \alpha & \beta \\ \beta & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = u_1^2\alpha + 2u_1u_2\beta + u_2^2 \geq u_1^2(\alpha - 2\beta^2) + \frac{1}{2}u_2^2$$

implying the matrix inequality

$$\begin{pmatrix} \alpha & \beta \\ \beta & 1 \end{pmatrix} \geq \begin{pmatrix} \alpha - 2\beta^2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}. \tag{47}$$

If $k \geq d^2$, then

$$\tau = \frac{R}{9\sqrt{d}}\Big(1 \wedge \frac{d\sigma}{\sqrt{k}R}\Big) \leq \frac{\sigma}{9}\sqrt{\frac{d}{k}} \leq \frac{\sigma}{4}. \tag{48}$$

The last inequality is rather loose but sufficient to give $8\tau^2 \leq \sigma^2/2$. Together with the matrix inequality applied to $\beta = -2\tau\langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle$, we obtain

$$S_1 \geq \begin{pmatrix} \frac{1}{2}\sigma^2 + \|\boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \geq \frac{1}{2}\begin{pmatrix} \sigma^2 & 0 \\ 0 & 1 \end{pmatrix} =: \Lambda_1.$$

Setting $\Lambda_0 := S_0$, we can now apply (44) for these choices of $S_0, S_1, \Lambda_0, \Lambda_1$. Observing that

$$\begin{aligned}\|\boldsymbol{\theta} - 2\tau\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2 - \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2 \\ = 4\tau^2 - 4\tau\langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle,\end{aligned}$$

$$S_1 - S_0 = \begin{pmatrix} 4\tau^2 - 4\tau\langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle & -2\tau\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle \\ -2\tau\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle & 0 \end{pmatrix},$$

and

$$\frac{1}{2}\operatorname{tr}\left(\begin{pmatrix} \lambda_0^{-1} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ \beta & 0 \end{pmatrix} 2 \begin{pmatrix} \lambda_1^{-1} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} \alpha & \beta \\ \beta & 0 \end{pmatrix}\right) \tag{49}$$

$$= \operatorname{tr}\left(\begin{pmatrix} \alpha\lambda_0^{-1} & \beta\lambda_0^{-1} \\ \beta & 0 \end{pmatrix}\begin{pmatrix} \alpha\lambda_1^{-1} & \beta\lambda_1^{-1} \\ \beta & 0 \end{pmatrix}\right) \tag{50}$$

$$= \operatorname{tr}\left(\begin{pmatrix} \alpha^2\lambda_0^{-1}\lambda_1^{-1} + \beta^2\lambda_0^{-1} & \alpha\beta\lambda_0^{-1}\lambda_1^{-1} \\ \alpha\beta\lambda_1^{-1} & \beta^2\lambda_1^{-1} \end{pmatrix}\right) \tag{51}$$

$$= \frac{\alpha^2}{\lambda_0\lambda_1} + \beta^2\Big(\frac{1}{\lambda_0} + \frac{1}{\lambda_1}\Big), \tag{52}$$

we find for the specific choices $\lambda_0 = \sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2$, $\lambda_1 = \sigma^2$, $\alpha^2 = (4\tau^2 - 4\tau\langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle)^2 \le 32\tau^4 + 32\tau^2\langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle^2$ and $\beta = -2\tau\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle$, using (44), and $\lambda_0 \ge \sigma^2$,

$$\begin{aligned}&\operatorname{KL}\left(P_{\boldsymbol{\theta},(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}, P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}\right) \\ &\le \frac{1}{2}\operatorname{tr}\left(\Lambda_0^{-1}(S_1 - S_0)\Lambda_1^{-1}(S_1 - S_0)\right) \\ &= \frac{\alpha^2}{\lambda_0\lambda_1} + \beta^2\Big(\frac{1}{\lambda_0} + \frac{1}{\lambda_1}\Big) \\ &\le \frac{32\tau^4 + 32\tau^2\langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle^2}{(\sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2)\sigma^2} + 6\frac{\tau^2}{\sigma^2}\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle^2 \\ &\le 32\frac{\tau^4}{\sigma^4} + 32\frac{\tau^2\langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle^2}{\sigma^2\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2} + 6\frac{\tau^2}{\sigma^2}\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle^2.\end{aligned}$$

Interchanging the sums over j and ℓ, we conclude from (42), (43), the previous inequality, $\sum_{j=1}^d \langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle^2 = \|\mathbf{v}'_\ell\|^2$, $\sum_{j=1}^d \langle \mathbf{e}_j, \boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\rangle^2 = \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2$, the

normalization $\|\mathbf{v}'_\ell\| = 1$, $\tau \le \sigma\sqrt{d/(81k)}$ (as derived in (48)), and $k \ge d^2$,

$$
\begin{aligned}
&\Big(\frac{1}{d}\sum_{j=1}^{d} \mathrm{TV}(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k})\Big)^2 \\
&\le \frac{1}{d}\sum_{j=1}^{d} \mathrm{TV}(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k})^2 \\
&\le \frac{1}{2^d d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d}\sum_{j=1}^{d} \mathrm{KL}\left(P_{\boldsymbol{\theta},\mathcal{V}_k}, P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,\mathcal{V}_k}\right)\mathbf{1}\big(\theta_j = \theta^\star_j + \tau\big) \\
&\le \frac{1}{2^d d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d}\sum_{j=1}^{d}\sum_{\ell=1}^{k} 32\frac{\tau^4}{\sigma^4} + 32\frac{\tau^2\langle \mathbf{e}_j, \boldsymbol{\theta}-\mathbf{v}_\ell - \langle\boldsymbol{\theta}-\mathbf{v}_\ell, \mathbf{v}'_\ell\rangle\mathbf{v}'_\ell\rangle^2}{\sigma^2\|\boldsymbol{\theta}-\mathbf{v}_\ell - \langle\boldsymbol{\theta}-\mathbf{v}_\ell,\mathbf{v}'_\ell\rangle\mathbf{v}'_\ell\|^2} + 6\frac{\tau^2}{\sigma^2}\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle^2 \\
&= \frac{1}{2^d d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d}\sum_{\ell=1}^{k} 32d\frac{\tau^4}{\sigma^4} + 38\frac{\tau^2}{\sigma^2} \\
&= 32k\frac{\tau^4}{\sigma^4} + 38\frac{\tau^2 k}{\sigma^2 d} \\
&\le \frac{32d^2}{81^2 k} + \frac{38}{81} \\
&\le \frac{1}{2}.
\end{aligned}
$$

Taking square roots, gives $1 - d^{-1}\sum_{j=1}^{d}\mathrm{TV}(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}) \ge 1 - 1/\sqrt{2}$. Since by definition $\tau = \frac{R}{9\sqrt{d}}(1 \wedge \frac{d\sigma}{\sqrt{k}R})$, the claimed lower bound follows from (39),

$$
\begin{aligned}
\sup_{\boldsymbol{\theta}\in\Theta} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\right] &\ge \frac{\tau^2 d}{2}\sup_{\boldsymbol{\theta}^\star\in B_{R/2}(0)}\Big(1 - \frac{1}{d}\sum_{j=1}^{d}\mathrm{TV}(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k})\Big) \\
&= \frac{1}{162}\Big(1 - \frac{1}{\sqrt{2}}\Big)\Big(R^2 \wedge \frac{d^2}{k}\sigma^2\Big).
\end{aligned}
$$

□

Proof of Theorem 4. Let

$$
\rho = c\frac{R}{\sqrt{d}}\Big(1 \wedge \frac{d}{\sqrt{k}}\Big(1 \vee \frac{\sigma}{R}\Big)\Big) \quad \text{for} \quad c = 2^{-8}. \tag{53}
$$

For any $\boldsymbol{\theta}^\star \in B_{R/2}(0)$ and any $\boldsymbol{\theta} \in \boldsymbol{\theta}^\star + \{-\rho,\rho\}^d$, we have $\|\boldsymbol{\theta}\| \le \|\boldsymbol{\theta}^\star\| + cR \le R$. Thus, we can lower bound $\sup_{\boldsymbol{\theta}\in B_R(0)} \ge \sup_{\boldsymbol{\theta}^\star\in B_{R/2}(0)} \max_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\rho,\rho\}^d}$. Lemma 6 gives

$$
\sup_{\boldsymbol{\theta}\in B_R(0)} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k}\left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2\right] \ge \sup_{\boldsymbol{\theta}^\star\in B_{R/2}(0)} \frac{\rho^2 d}{2}\Big(1 - \frac{1}{d}\sum_{j=1}^{d}\mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big) \tag{54}
$$

with $\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}$ as in (40) and (41) with τ replaced by ρ.

Since the query vectors are deterministic, (43) becomes now

$$
\mathrm{KL}\left(P_{\boldsymbol{\theta},\mathcal{V}_k}, P_{\boldsymbol{\theta}-2\rho\mathbf{e}_j,\mathcal{V}_k}\right) = \sum_{\ell=1}^{k}\mathrm{KL}\left(P_{\boldsymbol{\theta},(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}, P_{\boldsymbol{\theta}-2\rho\mathbf{e}_j,(Z_\ell,Z'_\ell)|(\mathbf{v}_\ell,\mathbf{v}'_\ell)}\right), \tag{55}
$$

where $\boldsymbol{\theta}$ is an element in the set $\boldsymbol{\theta}^\star + \{-\rho, \rho\}^d$ with $\theta_j = \theta_j^\star + \rho$.

Let $\mathbf{v}'$ satisfy $\|\mathbf{v}'\| = 1$ and let $\mathbf{w}, \mathbf{w}' \in \{-1, 0, 1\}^d$. We now prove that $\|\mathbf{u}\| \geq Rk^{-1/(d-1)}/4$ implies

$$\sigma^2 + \big\|\mathbf{u} + \rho\mathbf{w} + \rho\langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\big\|^2 - 8\rho^2 \geq \frac{\sigma^2 \vee \|\mathbf{u}\|^2}{8}. \tag{56}$$

To see this, use that $\|\mathbf{v}'\| = 1$ and $|\langle \mathbf{w}', \mathbf{v}'\rangle| \leq \|\mathbf{w}'\|$. Thus, triangle inequality yields $\rho\|\mathbf{w} + \langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\| \leq 2\rho\sqrt{d}$. By (53), $\rho = cRd^{-1/2}(1 \wedge dk^{-1/2}(1 \vee (\sigma/R))$.

If $k \leq 1/(16c)^{d-1}$, we have $\|\mathbf{u}\| \geq 4Rc \geq 4\rho\sqrt{d}$. Moreover, if $k > 1/(16c)^{d-1}$ and $\sigma \leq R$, we can write $k = (a/16c)^{d-1}$ with $a > 1$. Since $y \leq e^{y-1}$ and $c \leq 1/(16e^2)$, we have $d \leq (e^2)^{(d-1)/2} \leq 1/(16c)^{(d-1)/2}$. Recalling that $d \geq 3$, we find again $\|\mathbf{u}\| \geq Rk^{-1/(d-1)}/4 = 4cR/a \geq 4cRd/(a^{(d-1)/2}(1/16c)^{(d-1)/2}) = 4cRd/\sqrt{k} \geq 4\rho\sqrt{d}$. Thus in both of the previous cases $k \leq 1/(16c)^{d-1}$ and $\{k > 1/(16c)^{d-1}\} \cap \{\sigma \leq R\}$, $\|\mathbf{u}\| \geq 4\rho\sqrt{d} \geq 2\rho\|\mathbf{w} + \langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\|$ and therefore by triangle inequality, $\|\mathbf{u} + \rho\mathbf{w} + \rho\langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\| \geq \|\mathbf{u}\|/2$. Since for $d \geq 4$, also $\|\mathbf{u}\|^2/8 \geq 2\rho^2 d \geq 8\rho^2$, (56) follows for these cases.

The remaining case $k > 1/(16c)^{d-1}$ and $\sigma > R$, yields $\sigma/4 \geq 2cdk^{-1/2}\sigma \geq 2\rho\sqrt{d}$. Together with the elementary inequality $a^2 + b^2 \geq \frac{1}{2}(a+b)^2$, triangle inequality and $\rho\|\mathbf{w} + \langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\| \leq 2\rho\sqrt{d}$, we get $\sigma + \big\|\mathbf{u} + \rho\mathbf{w} + \rho\langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\big\| \geq \sigma + \big\|\mathbf{u}\big\| - 2\rho\sqrt{d} \geq (3/4)\sigma + \big\|\mathbf{u}\big\|$ and

$$\sigma^2 + \big\|\mathbf{u} + \rho\mathbf{w} + \rho\langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\big\|^2 \geq \frac{1}{2}\Big(\sigma + \big\|\mathbf{u} + \rho\mathbf{w} + \rho\langle \mathbf{w}', \mathbf{v}'\rangle \mathbf{v}'\big\|\Big)^2 \geq \frac{\sigma^2 \vee \|\mathbf{u}\|^2}{4}.$$

Moreover $\sigma^2/8 \geq 8\rho^2 d \geq 8\rho^2$. Thus, (56) holds in all cases.

Let

$$S_0 = \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2 & 0 \\ 0 & 1 \end{pmatrix}$$

be as in (45).

For the next arguments, we will assume that

$$\big\|\boldsymbol{\theta}^\star - \mathbf{v}_\ell + \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle\big\| \geq \frac{R}{4} k^{-1/(d-1)}. \tag{57}$$

Note that $\boldsymbol{\theta} \in \boldsymbol{\theta}^\star + \{-\rho, \rho\}^d$. Thus, applying (56) with $\mathbf{u} = \boldsymbol{\theta}^\star - \mathbf{v}_\ell + \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle$, $\rho\mathbf{w} = \boldsymbol{\theta} - \boldsymbol{\theta}^\star$, we have that $S_0 \geq \widetilde{\Lambda}_0$ with

$$\widetilde{\Lambda}_0 = \begin{pmatrix} \frac{\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2}{8} & 0 \\ 0 & 1 \end{pmatrix}.$$

Similarly, let

$$S_1 = \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - 2\rho\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2 & -2\rho\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle \\ -2\rho\langle \mathbf{e}_j, \mathbf{v}'_\ell\rangle & 1 \end{pmatrix}$$

be as in (46). Using the matrix inequality (47), $\langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle \le \|\mathbf{e}_j\| \|\mathbf{v}'_\ell\| = 1$, and applying (56) with $\mathbf{w} = \boldsymbol{\theta} - \boldsymbol{\theta}^\star - 2\rho \mathbf{e}_j$ and $\mathbf{w}' = \boldsymbol{\theta} - \boldsymbol{\theta}^\star$, we have

$$\begin{aligned} S_1 &\ge \begin{pmatrix} \sigma^2 + \|\boldsymbol{\theta} - 2\rho\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2 - 8\rho^2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix} \\ &\ge \begin{pmatrix} \frac{\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2}{8} & 0 \\ 0 & \frac{1}{2} \end{pmatrix}. \end{aligned}$$

Recall that $\boldsymbol{\theta}$ is an element in the set $\boldsymbol{\theta}^\star + \{-\rho, \rho\}^d$ with $\theta_j = \theta^\star_j + \rho$. This means that $\langle \boldsymbol{\theta} - \rho \mathbf{e}_j - \boldsymbol{\theta}^\star, \mathbf{e}_j \rangle = 0$ and thus

$$\begin{aligned} &\|\boldsymbol{\theta} - 2\rho\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2 - \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2 \\ &\qquad = -4\rho \langle \mathbf{e}_j, \boldsymbol{\theta} - \rho\mathbf{e}_j - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell \rangle \\ &\qquad = -4\rho \langle \mathbf{e}_j, \boldsymbol{\theta}^\star - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell \rangle. \end{aligned}$$

This implies

$$S_1 - S_0 = \begin{pmatrix} -4\rho \langle \mathbf{e}_j, \boldsymbol{\theta}^\star - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell \rangle & -2\rho \langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle \\ -2\rho \langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle & 0 \end{pmatrix} = \begin{pmatrix} \alpha & \beta \\ \beta & 0 \end{pmatrix}.$$

Combining (44) and (52) with $\lambda_0 = \lambda_1/2 = \frac{1}{8}(\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2)$, and α, β as above, we obtain

$$\begin{aligned} \mathrm{KL}\left(P_{\boldsymbol{\theta}, (Z_\ell, Z'_\ell) | (\mathbf{v}_\ell, \mathbf{v}'_\ell)}, P_{\boldsymbol{\theta} - 2\rho\mathbf{e}_j, (Z_\ell, Z'_\ell) | (\mathbf{v}_\ell, \mathbf{v}'_\ell)}\right) \le & 32 \frac{(-4\rho \langle \mathbf{e}_j, \boldsymbol{\theta}^\star - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell \rangle)^2}{(\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2)^2} \\ & + 12 \frac{(2\rho \langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle)^2}{\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2}. \end{aligned}$$

Since $\langle \boldsymbol{\theta} - \boldsymbol{\theta}^\star, \mathbf{v}'_\ell \rangle^2 \le \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|^2 \|\mathbf{v}'_\ell\|^2 = \rho^2 d$, we find using $(a+b)^2 \le 2a^2 + 2b^2$,

$$\langle \mathbf{e}_j, \boldsymbol{\theta}^\star - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell \rangle^2 \le 2 \langle \mathbf{e}_j, \boldsymbol{\theta}^\star - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell \rangle^2 + 2\rho^2 d \langle \mathbf{e}_j, \mathbf{v}'_\ell \rangle^2.$$

Since $\mathbf{e}_1, \ldots, \mathbf{e}_d$ form an orthonormal basis, $\sum_{j=1}^d \langle \mathbf{e}_j, \mathbf{a} \rangle^2 = \|\mathbf{a}\|^2$ for all d-dimensional vectors $\mathbf{a}$. Thus taking the sum over j yields

$$\begin{aligned} &\sum_{j=1}^d \mathrm{KL}\left(P_{\boldsymbol{\theta}, (Z_\ell, Z'_\ell) | (\mathbf{v}_\ell, \mathbf{v}'_\ell)}, P_{\boldsymbol{\theta} - 2\rho\mathbf{e}_j, (Z_\ell, Z'_\ell) | (\mathbf{v}_\ell, \mathbf{v}'_\ell)}\right) \\ &\le (2^8 + 48) \frac{\rho^2}{\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2} + 2^8 \frac{\rho^4 d}{(\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2)^2}. \end{aligned}$$

Recall that by (55), $\mathrm{KL}\left(P_{\boldsymbol{\theta}, \mathcal{V}_k}, P_{\boldsymbol{\theta} - 2\rho\mathbf{e}_j, \mathcal{V}_k}\right) = \sum_{\ell=1}^k \mathrm{KL}\left(P_{\boldsymbol{\theta}, (Z_\ell, Z'_\ell) | (\mathbf{v}_\ell, \mathbf{v}'_\ell)}, P_{\boldsymbol{\theta} - 2\rho\mathbf{e}_j, (Z_\ell, Z'_\ell) | (\mathbf{v}_\ell, \mathbf{v}'_\ell)}\right)$. Hence,

$$\begin{aligned} &\sum_{j=1}^d \mathrm{KL}\left(P_{\boldsymbol{\theta}, \mathcal{V}_k}, P_{\boldsymbol{\theta} - 2\rho\mathbf{e}_j, \mathcal{V}_k}\right) \\ &\le \sum_{\ell=1}^k \frac{176\rho^2}{\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2} + \sum_{\ell=1}^k \frac{2^8 \rho^4 d}{(\sigma^2 \vee \|\boldsymbol{\theta}^\star - \mathbf{v}_\ell - \langle \boldsymbol{\theta}^\star - \mathbf{v}_\ell, \mathbf{v}'_\ell \rangle \mathbf{v}'_\ell\|^2)^2}. \end{aligned}$$

The right hand side does not depend on $\boldsymbol{\theta}$ anymore. Combined with (42),

$$\begin{aligned}
&\Big(\frac{1}{d}\sum_{j=1}^{d}\mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k},\mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big)^2\\
&\le \frac{1}{2^d d}\sum_{\boldsymbol{\theta}\in\boldsymbol{\theta}^\star+\{-\tau,\tau\}^d}\sum_{j=1}^{d}\mathrm{KL}\left(P_{\boldsymbol{\theta},\mathcal{V}_k},P_{\boldsymbol{\theta}-2\tau\mathbf{e}_j,\mathcal{V}_k}\right)\mathbf{1}\big(\theta_j=\theta_j^\star+\tau\big) \qquad (58)\\
&\le \sum_{\ell=1}^{k}\frac{88\rho^2}{d(\sigma^2\vee\|\boldsymbol{\theta}^\star-\mathbf{v}_\ell-\langle\boldsymbol{\theta}^\star-\mathbf{v}_\ell,\mathbf{v}'_\ell\rangle\mathbf{v}'_\ell\|^2)}+\sum_{\ell=1}^{k}\frac{2^7\rho^4}{(\sigma^2\vee\|\boldsymbol{\theta}^\star-\mathbf{v}_\ell-\langle\boldsymbol{\theta}^\star-\mathbf{v}_\ell,\mathbf{v}'_\ell\rangle\mathbf{v}'_\ell\|^2)^2}.
\end{aligned}$$

Notice that this bound assumes the inequality (57).

If $\|\mathbf{v}'\|=1$, then $\|\mathbf{u}-\mathbf{v}-\langle\mathbf{u}-\mathbf{v},\mathbf{v}'\rangle\mathbf{v}'\|=\inf_{\gamma\in\mathbb{R}}\|\mathbf{u}-\mathbf{v}-\gamma\mathbf{v}'\|$. Thus, the set $T_{\mathbf{v},\mathbf{v}'}(r):=\{\mathbf{u}:\|\mathbf{u}-\mathbf{v}-\langle\mathbf{u}-\mathbf{v},\mathbf{v}'\rangle\mathbf{v}'\|\le r\}$ is a tube with radius $r>0$. The Euclidean ball with radius $r'>0$ and center $\mathbf{w}$ is denoted by $B_{r'}(\mathbf{w}):=\{\mathbf{y}:\|\mathbf{y}-\mathbf{w}\|\le r'\}$. If $\mathbf{u}\in B_{r'}(0)$, then $\langle-\mathbf{v},\mathbf{v}'\rangle-r'\le\langle\mathbf{u}-\mathbf{v},\mathbf{v}'\rangle\le\langle-\mathbf{v},\mathbf{v}'\rangle+r'$. Thus,

$$H_{\mathbf{v},\mathbf{v}'}(r,r'):=T_{\mathbf{v},\mathbf{v}'}(r)\cap B_{r'}(0)\subseteq\Big\{\mathbf{u}:\inf_{\langle-\mathbf{v},\mathbf{v}'\rangle-r'\le\gamma\le\langle-\mathbf{v},\mathbf{v}'\rangle+r'}\|\mathbf{u}-\mathbf{v}-\gamma\mathbf{v}'\|\le r\Big\}.$$

Let $\Gamma(\cdot)$ be the Gamma function. The volume formula for a tube (e.g. (2.3) in [17]) says that the volume of the right hand side is bounded by $2r'c_{d-1}r^{d-1}+c_dr^d$ with $c_p:=\pi^{p/2}/\Gamma(p/2+1)$ the volume of a p-dimensional unit ball. By (41) in [23], $\Gamma(d/2+1)/\Gamma(d/2+1/2)\le\sqrt{d/2+1/2}$. Therefore $c_{d-1}\le\sqrt{\pi d}c_d$. Moreover, $d\ge5$ implies $1+2\sqrt{\pi}\sqrt{d/2+1/2}\le2^{d-2}$. Thus, $2c_{d-1}+c_d\le(1+2\sqrt{\pi}\sqrt{d/2+1/2})c_d\le2^{d-2}c_d$. Together with the previous display, this means that

$$\mathrm{Vol}\left(H_{\mathbf{v},\mathbf{v}'}(r,r')\right)\le2^{d-2}r'c_dr^{d-1},\quad\text{whenever}\quad r\le r'.$$

with Vol the volume (Lebesgue measure) in d-dimensions. Using again that $d\ge4$, this means that for any $\mathbf{v}_1,\mathbf{v}'_1,\dots,\mathbf{v}_k,\mathbf{v}'_k$, and any $b>0$,

$$\begin{aligned}
\mathrm{Vol}\left(H_{\mathbf{v}_\ell,\mathbf{v}'_\ell}\Big(\frac{bR}{4}k^{-1/(d-1)},\frac{R}{2}\Big)\right)&\le2^{d-3}Rc_db^{d-1}R^{d-1}k^{-1}4^{1-d}\\
&=\frac{b^{d-1}}{2k}\Big(\frac{R}{2}\Big)^dc_d\\
&=\frac{b^{d-1}}{2k}\mathrm{Vol}\left(B_{R/2}(0)\right)
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{Vol}\Big(\bigcup_{\ell=1}^{k}H_{\mathbf{v}_\ell,\mathbf{v}'_\ell}\Big(\frac{R}{4}k^{-1/(d-1)},\frac{R}{2}\Big)\Big)&\le\sum_{\ell=1}^{k}\mathrm{Vol}\left(H_{\mathbf{v}_\ell,\mathbf{v}'_\ell}\Big(\frac{R}{4}k^{-1/(d-1)},\frac{R}{2}\Big)\right)\\
&\le\frac{1}{2}\mathrm{Vol}\left(B_{R/2}(0)\right).
\end{aligned}$$

The latter implies that

$$\begin{aligned}
\mathrm{Vol}\Big(B_{R/2}(0)\setminus\bigcup_{\ell=1}^{k}H_{\mathbf{v}_\ell,\mathbf{v}'_\ell}\Big(\frac{R}{8}k^{-1/(d-1)},\frac{R}{2}\Big)\Big)&\ge\mathrm{Vol}\left(B_{R/2}(0)\right)-\frac{1}{2}\mathrm{Vol}\left(B_{R/2}(0)\right)\\
&=\frac{1}{2}\mathrm{Vol}\left(B_{R/2}(0)\right).
\end{aligned}$$

Now we define a probability measure on the ball $B_{R/2}(0)$ by

$$\nu(A) = \frac{\mathrm{Vol}\left(A \cap \left(B_{R/2}(0) \setminus \cup_{\ell=1}^{k} H_{\mathbf{v}_\ell, \mathbf{v}'_\ell}\left(Rk^{-1/(d-1)}/4, R/2\right)\right)\right)}{\mathrm{Vol}\left(B_{R/2}(0) \setminus \cup_{\ell=1}^{k} H_{\mathbf{v}_\ell, \mathbf{v}'_\ell}\left(Rk^{-1/(d-1)}/4, R/2\right)\right)}.$$

The distribution ensures that the inequality (57) holds for all $\ell = 1, \dots, k$ with probability one. Thus, we can apply (58) and therefore lower bound $\inf_{\boldsymbol{\theta}^\star \in B_{R/2}(0)}$ by an average with respect to the probability measure ν,

$$\begin{aligned}
&\inf_{\boldsymbol{\theta}^\star \in B_{R/2}(0)} \Big(\frac{1}{d}\sum_{j=1}^{d} \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k}\right)\Big)^2 \\
&\le \int \Big(\sum_{\ell=1}^{k} \frac{88\rho^2}{d(\sigma^2 \vee \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2)} + \sum_{\ell=1}^{k} \frac{2^7\rho^4}{(\sigma^2 \vee \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2)^2} \Big) d\nu(\boldsymbol{\theta}) \\
&\le \sum_{\ell=1}^{k} \frac{88\rho^2}{d} \Big(\frac{1}{\sigma^2} \wedge \int \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2} \Big) d\nu(\boldsymbol{\theta}) \\
&\quad + 2^7\rho^4 \Big(\frac{1}{\sigma^4} \wedge \int \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^4} \Big) d\nu(\boldsymbol{\theta}). \qquad (59)
\end{aligned}$$

The sum $\sum_{0 \le t \le \log_2(k^{1/(d-1)})-1}$ is set to zero if $k < 2^{d-1}$. Using that for $d \ge 6$, one has $2^{d-5} - 1 \ge \frac{1}{2}2^{d-5}$, we obtain $\sum_{t=0}^{q} 2^{t(d-5)} = (2^{(q+1)(d-5)} - 1)/(2^{d-5} - 1) \le 2 \cdot 2^{q(d-5)}$. Applying this with $q = \lfloor \log_2(k^{1/(d-1)}) - 1 \rfloor$ and using that $k^{-4/(d-1)}k = k^{(d-5)/(d-1)}$ yields

$$\begin{aligned}
&\int \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^4} d\nu(\boldsymbol{\theta}) \\
&\le \int_{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2 \ge R/8}^{\infty} \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^4} \, d\nu(\boldsymbol{\theta}) \\
&\quad + \frac{2}{\mathrm{Vol}\left(B_{R/2}(0)\right)} \sum_{0 \le t \le \log_2(k^{1/(d-1)})-1} \\
&\qquad \int_{2^t Rk^{-1/(d-1)}/4 \le \|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2 \le 2^{t+1} Rk^{-1/(d-1)}/4} \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^4} \, d\nu(\boldsymbol{\theta}) \\
&\le \frac{64}{R^4} + \frac{2}{\mathrm{Vol}\left(B_{R/2}(0)\right)} \sum_{0 \le t \le \log_2(k^{1/(d-1)})-1} \frac{\mathrm{Vol}\left(H_{\mathbf{v}_\ell, \mathbf{v}'_\ell}\left(2^{t+1}Rk^{-1/(d-1)}/4, R/2\right)\right)}{(2^{2t}R^2k^{-2/(d-1)}/16)^2} \\
&= \frac{64}{R^4} + \frac{2}{\mathrm{Vol}\left(B_{R/2}(0)\right)} \sum_{0 \le t \le \log_2(k^{1/(d-1)})-1} \frac{2^8}{2^{4t}R^4k^{-4/(d-1)}} \frac{2^{(t+1)(d-1)}}{2k} \mathrm{Vol}\left(B_{R/2}(0)\right) \\
&\le \frac{64}{R^4} + \frac{2^8 \cdot 2^5}{R^4} \\
&\le \frac{2^{14}}{R^4}. \qquad (60)
\end{aligned}$$

Since ν is a probability measure, Jensen's inequality gives

$$\int \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^2} d\nu(\boldsymbol{\theta}) \le \Big(\int \frac{1}{\|\boldsymbol{\theta} - \mathbf{v}_\ell - \langle \boldsymbol{\theta} - \mathbf{v}_\ell, \mathbf{v}'_\ell\rangle \mathbf{v}'_\ell\|^4} d\nu(\boldsymbol{\theta}) \Big)^{1/2} \le \frac{2^7}{R^2}.$$

Recall that $\rho = cRd^{-1/2}(1 \wedge dk^{-1/2}(1 \vee \sigma/R))$. Thus, $\rho^2 \le c^2 dk^{-1}(R \vee \sigma)^2$, $\rho^2 \le c^2R^2d^{-1}$, and

$\rho^4 \le \rho^2\rho^2 \le c^4 k^{-1}(R \vee \sigma)^2 R^2$. Combined with (59), (60), and $c = 2^{-8}$,

$$\inf_{\boldsymbol{\theta}^\star \in B_{R/2}(0)} \Big(\frac{1}{d} \sum_{j=1}^{d} \mathrm{TV}\left(\mathbb{P}_{+j,\boldsymbol{\theta}^\star,\mathcal{V}_k}, \mathbb{P}_{-j,\boldsymbol{\theta}^\star,\mathcal{V}_k} \right) \Big)^2 \le \frac{88 \cdot 2^7 k\rho^2}{d(\sigma^2 \vee R^2)} + \frac{2^{21} k\rho^4}{\sigma^4 \vee R^4} \le 88 \cdot 2^7 c^2 + 2^{21} c^4 \le \frac{1}{4}.$$

Using (54), we find that

$$\sup_{\boldsymbol{\theta} \in B_R(0)} \mathbb{E}_{\boldsymbol{\theta},\mathcal{V}_k} \left[\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right] \ge \frac{\rho^2 d}{4} = 2^{-18} R^2 \Big(1 \wedge \frac{d^2}{k} (R \vee \sigma)^2 \Big).$$

□

Proof of Corollary 1. Using $\boldsymbol{\theta}_0 = 0$ and Lemma 4 with $\mathbf{a} = \mathbf{b} = \boldsymbol{\theta}^\star$, we obtain that $\|S_0\| = \|\boldsymbol{\theta}^\star(\boldsymbol{\theta}^\star)^\intercal\| = \|\boldsymbol{\theta}^\star\|^2 \le d$. Thus Theorem 1 combined with (14) for $\kappa = 1$ and $\gamma = 2$ yields the rate $d^2 \log^2(d)/k$ for the upper bound in the adaptive query setting.

The minimax rate in the non-adaptive query setting is $R^2 \wedge \frac{d^2}{k}(R \vee \sigma)^2$. Using that $R = \sqrt{d} \ge \sigma$, this becomes $d^2 \wedge d^3/k$. Since $k \gtrsim d^2 \log(d)$ the second term always dominates and the minimax rate is d^3/k. □

Acknowledgement

We are grateful to Chao Gao for fruitful discussions. The research has been supported by the NWO Vidi grant VI.Vidi.192.021.

References

[1] Agarwal, A., Dekel, O., and Xiao, L. Optimal algorithms for online convex optimization with multi-point bandit feedback. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)* (2010).

[2] Arias-Castro, E., Candes, E. J., and Davenport, M. A. On the fundamental limits of adaptive sensing. *IEEE Transactions on Information Theory 59*, 1 (2013), 472–481.

[3] Bach, F., and Moulines, E. Non-strongly-convex smooth stochastic approximation with convergence rate O (1/n). *Advances in neural information processing systems 26* (2013).

[4] Bach, F., and Perchet, V. Highly-smooth zero-th order online optimization. In *Conference on Learning Theory* (2016), PMLR, pp. 257–283.

[5] Balasubramanian, K., and Ghadimi, S. Zeroth-order nonconvex stochastic optimization: handling constraints, high dimensionality, and saddle points. *Found. Comput. Math. 22*, 1 (2022), 35–76.

[6] Baydin, A. G., Pearlmutter, B. A., Syme, D., Wood, F., and Torr, P. Gradients without backpropagation. *arXiv preprint arXiv:2202.08587* (2022).

[7] Bello, K., and Honorio, J. Computationally and statistically efficient learning of causal Bayes nets using path queries. In *Advances in Neural Information Processing Systems* (2018), S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., vol. 31, Curran Associates, Inc.

[8] Bos, T., and Schmidt-Hieber, J. Convergence guarantees for forward gradient descent in the linear regression model. *arXiv preprint* (2023).

[9] Clara, G., Langer, S., and Schmidt-Hieber, J. Dropout regularization versus ℓ_2-penalization in the linear model. *arXiv preprint* (2023), arXiv:2306.10529.

[10] Crick, F. The recent excitement about neural networks. *Nature 337*, 6203 (1989), 129–132.

[11] Dieuleveut, A., Flammarion, N., and Bach, F. Harder, better, faster, stronger convergence rates for least-squares regression. *The Journal of Machine Learning Research 18*, 1 (2017), 3520–3570.

[12] Duchi, J. C., Jordan, M. I., Wainwright, M. J., and Wibisono, A. Optimal rates for zero-order convex optimization: The power of two function evaluations. *IEEE Transactions on Information Theory 61*, 5 (2015), 2788–2806.

[13] Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. *arXiv preprint cs/0408007* (2004).

[14] Frémaux, N., Sprekeler, H., and Gerstner, W. Functional requirements for reward-modulated spike-timing-dependent plasticity. *Journal of Neuroscience 30*, 40 (2010), 13326–13337.

[15] Hazan, E., Koren, T., and Levy, K. Y. Logistic regression: Tight bounds for stochastic and online optimization. In *Conference on Learning Theory* (2014), PMLR, pp. 197–209.

[16] Jin, Y., Xiao, T., and Balasubramanian, K. Statistical inference for Polyak-Ruppert averaged zeroth-order stochastic gradient algorithm. *arXiv preprint* (2021), arXiv:2102.05198.

[17] Johnstone, I., and Siegmund, D. On Hotelling's formula for the volume of tubes and Naiman's inequality. *Ann. Statist. 17*, 1 (1989), 184–194.

[18] Lakshminarayanan, C., and Szepesvari, C. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (09–11 Apr 2018), A. Storkey and F. Perez-Cruz, Eds., vol. 84 of *Proceedings of Machine Learning Research*, PMLR, pp. 1347–1355.

[19] Larson, J., Menickelly, M., and Wild, S. M. Derivative-free optimization methods. *Acta Numer. 28* (2019), 287–404.

[20] Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., and Hinton, G. Backpropagation and the brain. *Nature Reviews Neuroscience 21*, 6 (2020), 335–346.

[21] Lugosi, G., Truszkowski, J., Velona, V., and Zwiernik, P. Learning partial correlation graphs and graphical models by covariance queries. *Journal of Machine Learning Research 22*, 203 (2021), 1–41.

[22] Novitskii, V., and Gasnikov, A. Improved exploitation of higher order smoothness in derivative-free optimization. *Optimization Letters 16*, 7 (2022), 2059–2071.

[23] Qi, F., and Luo, Q.-M. Bounds for the ratio of two gamma functions: from Wendel's asymptotic relation to Elezović-Giordano-Pečarić's theorem. *J. Inequal. Appl.* (2013), 2013:542, 20.

[24] Regis, M., Serra, P., and van den Heuvel, E. R. Random autoregressive models: a structured overview. *Econometric Reviews 41*, 2 (2022), 207–230.

[25] Ren, M., Kornblith, S., Liao, R., and Hinton, G. Scaling forward gradient with local losses. *arXiv preprint arXiv:2210.03310* (2022).

[26] Schmidt-Hieber, J. Interpreting learning in biological neural networks as zero-order optimization method. *arXiv preprint* (2023), arXiv:2301.11777.

[27] Shamir, O. On the complexity of bandit and derivative-free stochastic convex optimization. In *Proceedings of the 26th Annual Conference on Learning Theory* (Princeton, NJ, USA, 2013), S. Shalev-Shwartz and I. Steinwart, Eds., vol. 30 of *Proceedings of Machine Learning Research*, PMLR, pp. 3–24.

[28] Steinhardt, J., Valiant, G., and Wager, S. Memory, communication, and statistical queries. In *29th Annual Conference on Learning Theory* (Columbia University, New York, New York, USA, 23–26 Jun 2016), V. Feldman, A. Rakhlin, and O. Shamir, Eds., vol. 49 of *Proceedings of Machine Learning Research*, PMLR, pp. 1490–1516.

[29] Trappenberg, T. P. *Fundamentals of Computational Neuroscience: Third Edition*. Oxford University Press, 12 2022.

[30] Tsybakov, A. B. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[31] Zhang, F. *Matrix theory*. Universitext. Springer-Verlag, New York, 1999. Basic results and techniques.