Statistics and Computing (2021) 31:50
https://doi.org/10.1007/s11222-021-10011-z

Control variate selection for Monte Carlo integration

Rémi Leluc[1] · François Portier[1] · Johan Segers[2]

Received: 4 July 2020 / Accepted: 27 March 2021 / Published online: 25 June 2021

Abstract
Monte Carlo integration with variance reduction by means of control variates can be implemented by the ordinary least squares estimator for the intercept in a multiple linear regression model with the integrand as response and the control variates as covariates. Even without special knowledge on the integrand, significant efficiency gains can be obtained if the control variate space is sufficiently large. Incorporating a large number of control variates in the ordinary least squares procedure may however result in (i) a certain instability of the ordinary least squares estimator and (ii) a possibly prohibitive computation time. Regularizing the ordinary least squares estimator by preselecting appropriate control variates via the Lasso turns out to increase the accuracy without additional computational cost. The findings in the numerical experiment are confirmed by concentration inequalities for the integration error.

Keywords Control variate · Lasso · Monte Carlo · Variable selection · Variance reduction

1 Introduction

Whereas the basic Monte Carlo (MC) estimate of an integral or expectation is given by $(1/n)\sum_i f_i$, for independent and identically distributed random variables f_i, the control variates method is based on $(1/n)\sum_i (f_i + h_i)$, where the variables h_i, called control variates, are constructed to have zero expectation. When the controls h_i have been selected or estimated properly (based on the samples f_i), the use of control variates might reduce the variance of the basic MC estimate significantly. The method of control variates, already used frequently to compute prices of financial derivatives (Glasserman 2013; Gobet and Labart 2010), has been employed recently in many different fields of Machine Learning and Statistics. Examples include (1) *reinforcement learning* and more particularly *policy gradient* methods (Jie and Abbeel 2010; Liu et al. 2018) where the score function permits to define many control variates; (2) inference in complex probabilistic models (Ranganath et al. 2014) where the Stein method allows to define accurate control variates (see e.g., Oates et al. 2017; Brosse et al. 2018; Belomestny et al. 2020 and the references therein); (3) gradient based *optimization* (Wang et al. 2013; Gower et al. 2018), (4) *time series analysis* when approximating the characteristic function (Davis et al. 2019), and (5) semi-supervised inference (Zhang et al. 2019).

Suppose that $m \geq 1$ control variates are available and $n \geq 1$ samples have been generated. Any linear combination of control variates can be used as a particular control variate. In terms of the variance of the estimation error, the optimal linear combination can be estimated based on the empirical risk minimization principle applied to an ordinary least squares (OLS) regression problem [see Eq. (3) below]. This approach, referred to as OLS, is the most common implementation of the control variates method as detailed for instance in (Owen 2013, Sect. 8.3) or (Portier and Segers 2019; South et al. 2018), although other implementations are possible, see Remark 2 below.

Asymptotically, the OLS error is bounded by the MC error and is proportional to the L_2 approximation error of the integrand in the linear span of control variates (Glynn and Szechtman 2002). In combination with well-known approximation results in L_p-spaces (Rudin 2006), this representation of the OLS error suggests to use an increasing number of con-

✉ Rémi Leluc
remi.leluc@gmail.com

François Portier
francois.portier@gmail.com

Johan Segers
johan.segers@uclouvain.be

[1] Télécom Paris, Institut Polytechnique de Paris, 19 Place Marguerite Perey, F-91120 Palaiseau, France

[2] UCLouvain, LIDAM/ISBA, Voie du Roman Pays 20/L1.04.01, 1348 Louvain-la-Neuve, Belgium

trol variates. Indeed, in Portier and Segers (2019) it is shown that when m grows with n, the OLS error rate can be faster than $1/\sqrt{n}$.

However, when based on a large number of control variates, the OLS suffers from two classical problems common for least squares methods: (1) numerical instabilities when the control variates are nearly collinear, and (2) a computational complexity in $m^3 + nm^2$, which might be prohibitive.

To deal with these two issues, it has been proposed in South et al. (2018) to regularize the OLS estimate by adding a ℓ_1-penalty term in the minimization problem, just as in the LASSO (Tibshirani 1996). Simulation results in South et al. (2018) show that this approach, referred to as LASSO, provides great improvements in practice. However, those practical findings are not supported by an asymptotic error rate nor by a non-asymptotic error bound.

The main objective of the paper is to provide a non-asymptotic theory for the use of control variates in Monte Carlo simulations. The contributions are as follows.

1. A *new method* called LSLASSO is proposed. In the spirit of Belloni and Chernozhukov (2013), it consists in selecting the best control variates via the LASSO, using subsampling to decrease the computation time, and then to apply OLS with the selected controls.
2. *Support recovery*: the LASSO is shown to select the correct control variates with large probability.
3. *Concentration inequalities* are derived for the OLS, LASSO and LSLASSO integration errors. The one for the OLS highlights a compromise between the approximation error of the integrand in the linear span of control variates and the multicollinearities between the control variates. The ones for (LS)LASSO show significant improvements regarding the effects of multicollinearity.

The approach for the proofs combines well known sub-Gaussian concentration inequalities (Boucheron et al. 2013) along with a lower bound for the smallest eigenvalue of an empirical Gram matrix, based on a Chernoff inequality for matrices (Tropp 2015, Theorem 5.1.1).

The outline of the paper is as follows. Section 2 introduces the theoretical background and the different MC estimates and provides some comments about their practical implementation and some possible alternative approaches. Section 3 contains the statements of the theoretical results. Sections 4 and 5 describe numerical experiments on artificial and real data to illustrate the finite-sample behavior of the methods. Section 6 concludes the main part of the paper with a discussion of avenues for further research. Section A contains some auxiliary results, whereas the proofs of the four theorems stated in Sect. 3 are given in Sects. B to E.

2 Monte Carlo integration and control variates

Background. Let $f \in L_2(P)$ be a square integrable, real-valued function on a probability space $(\mathscr{X}, \mathscr{A}, P)$ of which we would like to calculate the integral

$$P(f) = \int_{\mathscr{X}} f(x)\, P(dx).$$

The MC estimator of $P(f)$ based on independent random variables $X_1, \ldots, X_n$ taking values in $\mathscr{X}$ and with common distribution P is

$$\hat{\alpha}_n^{\mathrm{mc}}(f) = P_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

This estimator is unbiased and has variance $n^{-1}\sigma_0^2(f)$, where $\sigma_0^2(f) = P[(f - P(f))^2]$.

The control variates are functions $h_1, \ldots, h_m \in L_2(P)$ with known expectations. Without loss of generality, assume that $P(h_k) = 0$ for all $k \in \{1, \ldots, m\}$. Let $h = (h_1, \ldots, h_m)^T$ denote the $\mathbb{R}^m$-valued function with the m control variates as elements. Let $\mathscr{F}_m = \mathrm{Span}\{h_1, \ldots, h_m\} = \{\beta^T h : \beta \in \mathbb{R}^m\}$ denote the closed linear subspace of $L_2(P)$ generated by the control variates.

For any coefficient vector $\beta = (\beta_1, \ldots, \beta_m)^T \in \mathbb{R}^m$, we have $P(f - \beta^T h) = P(f)$, so that $P_n(f - \beta^T h)$ is an unbiased estimator of $P(f)$, with variance $n^{-1}P[(f - P(f) - \beta^T h)^2]$. Any oracle coefficient

$$\beta^\star(f) \in \operatorname*{arg\,min}_{\beta \in \mathbb{R}^m} P[(f - P(f) - \beta^T h)^2]$$

minimizes the variance. If such a $\beta^\star(f)$ would be known, the resulting oracle estimator would be

$$\hat{\alpha}_n^{\mathrm{or}}(f) = P_n[f - \beta^\star(f)^T h]. \tag{1}$$

By definition, the oracle estimator achieves the minimal variance $n^{-1}\sigma_m^2(f)$ where $\sigma_m^2(f)$ is the minimum value of the variance term $P[(f - P(f) - \beta^T h)^2]$ with respect to β. For any $m' \in \{0, 1, \ldots, m\}$, if we use only the first m' control variates $h_1, \ldots, h_{m'}$, or even none at all in case $m' = 0$, we have $\sigma_m^2(f) \le \sigma_{m'}^2(f)$. In particular, if $\beta^\star(f)$ would be known, the use of control variates would always reduce the variance of the basic Monte Carlo estimator.

As $\beta^\star(f)^T h$ is the $L_2(P)$-projection of $f - P(f)$ on the linear vector space $\mathscr{F}_m$ and since the control variates are centered, $\beta^\star(f)$ satisfies the normal equations $P(hh^T)\beta^\star(f) = P(hf)$. The integral $P(f)$ thus appears as the intercept of a linear regression model with response f and explanatory

variables $h_1, \ldots, h_m$, and it can be expressed as

$$(P(f), \beta^\star(f)) \in \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} P[(f-\alpha-\beta^T h)^2]. \tag{2}$$

The empirical risk minimization paradigm applied to the risk function on the right-hand side of (2) will lead to the OLS and LASSO estimates, to be defined further in this section. The same paradigm suggests the use of other regression methods for MC integration such as Principal Component Regression (PCR) or Ridge Regression, which will not be considered in this paper.

Remark 1 **(Choice of control variates)** Which control variates work well depends on the problem. In the Black–Scholes model, for instance, an effective control variate for the price of an option is the geometric average of the price series (Glasserman 2013, Example 4.1.2). Two generic ways to construct control variates are to be noted. Whenever $P(dx) = w(x)\,Q(dx)$, where $w : \mathscr{X} \to [0,\infty)$ and Q is a probability measure on $(\mathscr{X}, \mathscr{A})$, the quantity of interest is $P(f) = Q(wf)$, so that we can use control variates for wf with respect to Q. This trick can be useful in combination with importance sampling (Owen and Zhou 2000). If P has density p with respect to the Lebesgue measure and if we have access to the derivatives of p, Stein's method might be used to build infinitely many control functions (Oates et al. 2017).

Ordinary Least Squares Monte Carlo. Replacing the distribution P by the sample measure P_n in (2), we obtain the OLS estimator $\hat{\alpha}_n^{\mathrm{ols}}(f)$ of $P(f)$ as a minimizer of the empirical risk

$$\mathscr{R}_n(\alpha,\beta) = \left\| f^{(n)} - \alpha \mathbb{1}_n - H\beta \right\|_2^2$$

given by

$$\left(\hat{\alpha}_n^{\mathrm{ols}}(f), \hat{\beta}_n^{\mathrm{ols}}(f)\right) \in \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \mathscr{R}_n(\alpha,\beta) \tag{3}$$

where $\|\cdot\|_2$ denotes the Euclidean norm, $\mathbb{1}_n = (1,\ldots,1)^T \in \mathbb{R}^n$, $f^{(n)} = (f(X_1),\ldots,f(X_n))^T \in \mathbb{R}^n$ and H is the random $n \times m$ matrix defined by

$$H = \left(h_j(X_i)\right)_{\substack{i=1,\ldots,n \\ j=1,\ldots,m}}.$$

The minimization problem in (3) can be expressed using an OLS estimate with centered variables as

$$\begin{cases} \hat{\alpha}_n^{\mathrm{ols}}(f) = P_n[f - \hat{\beta}_n^{\mathrm{ols}}(f)^T h], \\ \hat{\beta}_n^{\mathrm{ols}}(f) \in \underset{\beta\in\mathbb{R}^m}{\arg\min} \left\| f_c^{(n)} - H_c\beta \right\|_2^2, \end{cases} \tag{4}$$

where $f_c^{(n)} = f^{(n)} - \mathbb{1}_n(\mathbb{1}_n^T f^{(n)})/n$ and $H_c = H - \mathbb{1}_n(\mathbb{1}_n^T H)/n$. Indeed, for fixed $\beta \in \mathbb{R}^m$, the minimizer over $\alpha \in \mathbb{R}$ of the objective function in (3) is just $P_n(f-\beta^T h) = P_n(f) - \beta^T P_n(h)$, and since $P_n(f) = (\mathbb{1}_n^T f^{(n)})/n$ and $P_n(h) = (\mathbb{1}_n^T H)/n$, the equivalence of (3) and (4) follows.

Remark 2 **(Variations)** The solution of the linear regression problem (4) involves the empirical covariance matrix defined by $n^{-1}H_c^T H_c = P_n(hh^T) - P_n(h)P_n(h^T)$. Using different estimates of the Gram matrix $P(hh^T)$ leads to alternative control variate MC estimates for $P(f)$ (Glynn and Szechtman 2002; Portier and Segers 2019). For fixed m and as $n \to \infty$, all these estimators are consistent and asymptotically normal. The OLS estimator, however, is the only one that can integrate both the constant functions and the control functions without error.

Remark 3 **(Invariance)** The OLS estimator does not change if we replace the control variate vector h by Ah, where A is an arbitrary invertible $m \times m$ matrix. Provided the control functions are linearly independent, the property of isotropy, i.e., $P(hh^T) = I_m$, can therefore always be enforced by an appropriate linear transformation of the vector of control variates.

Remark 4 **(Computation time)** The computation time of the OLS method is of the order $nm^2 + m^3 + nt$, where nm^2 and m^3 operations are needed for computing and inverting $H_c^T H_c$ respectively and where t stands for the time needed to evaluate f. Computational benefits occur when there are multiple integrands, since the OLS estimate can be represented as $w^T f^{(n)}$, where the weight vector $w \in \mathbb{R}^n$ does not depend on the integrands (Portier and Segers 2019). If q integrals need to be evaluated, the computing time becomes $nm^2 + m^3 + qnt$, since the matrix $H_c^T H_c$ only depends on the control variates but not on the integrand.

Remark 5 **(Variance reduction)** The advantage of using a given set of m control variates over standard MC can be assessed through the value of the residual standard deviation $\sigma_m(f)$. In Portier and Segers (2019), bounds for $\sigma_m(f)$ are computed in specific examples. For instance, if $\mathscr{X} = [-1,1]^d$ and the h_k are tensor products of Legendre polynomials, then for any k-times continuously differentiable function f it holds that $\sigma_m(f) = O(m^{-k/d})$ as $m \to \infty$. This bound emphasizes the benefits of using polynomials when the integrand is regular.

LASSO Monte Carlo. The LASSO, introduced in Tibshirani (1996), is a regression technique that consists in minimizing the usual least squares loss plus an ℓ_1-penalty term on the vector of regression coefficients. In contrast with OLS, the LASSO usually produces a vector with many zero coefficients, meaning that the corresponding variables are no longer included in the predictive model. The LASSO thus

achieves estimation and variable selection at the same time. As the use of control variates in MC integration is linked with regression, the LASSO can take advantage from situations where many control variates are present but not all of them are useful.

The LASSO estimator $\hat{\alpha}_n^{\text{lasso}}(f)$ of $P(f)$ follows from adding a ℓ_1-penalization to the objective function in (3). It is formally defined as

$$\left(\hat{\alpha}_n^{\text{lasso}}(f), \hat{\beta}_n^{\text{lasso}}(f)\right) \in \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^m}{\arg\min} \frac{1}{2n}\mathscr{R}_n(\alpha,\beta) + \lambda \|\beta\|_1$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm on Euclidean space. By the same argument used to justify the equivalence of (3) and (4), the LASSO can be based on centered variables via

$$\begin{cases} \hat{\alpha}_n^{\text{lasso}}(f) = P_n[f - \hat{\beta}_n^{\text{lasso}}(f)^T h], \\ \hat{\beta}_n^{\text{lasso}}(f) \in \underset{\beta\in\mathbb{R}^m}{\arg\min} \dfrac{1}{2n}\left\| f_c^{(n)} - H_c\beta \right\|_2^2 + \lambda \|\beta\|_1 . \end{cases} \tag{5}$$

Remark 6 **(Computation)** For the practical implementation of the LASSO, it is commonly recommended to first center and rescale the explanatory variables empirically (Tibshirani et al. 2015, Sect. 2.2). The centering by the sample mean is taken care of in (5). However, for ease of presentation, no empirical rescaling of the control variates is considered in the theoretical analysis. This is in line with the approach proposed in (Tibshirani et al. 2015, Chapter 11). Still, such rescaling is done in the simulation experiments reported in Sect. 4.

Remark 7 **(Computation time)** The LASSO solution is usually computed approximately by *cyclical coordinate descent*. At each iteration, this algorithm minimizes (5) with respect to a single coordinate, say β_k, while considering other coordinates, $\beta_{(-k)} \in \mathbb{R}^{m-1}$, as constant. This one-dimensional optimization problem has an explicit argmin. Let $H_{c,k}$ be the k-th column of H_c that has been normalized such that $\|H_{c,k}\|_2 = 1$ (as indicated in the previous remark). The argmin is then simply given by $\eta_\lambda(\langle z_k, H_{c,k}\rangle)$ where $z_k = f_c^{(n)} - H_{c,(-k)}\beta_{(-k)}$, $H_{c,(-k)}$ is obtained by removing $H_{c,k}$ from H_c and η is the *soft-thresholding function* (Tibshirani et al. 2015, Sect. 2.4, Eq. (2.14)). Since n operations are needed to update z_k and the same number is needed to compute the scalar product, the LASSO requires only $nD + nt$ operations, where D stands for the number of iterations conducted in the cyclical coordinate descent and t represents the time needed to evaluate f. The value of D is often imposed by a stopping rule within the algorithm but it could also be fixed by the user in order to control the computing time. The selection of the next coordinate k to update can be done cyclically or at random.

LSLASSO Monte Carlo. The application of ordinary least squares after model selection by the LASSO has been recently studied in Belloni and Chernozhukov (2013). They show, in the setting of nonparametric regression, that OLS post-LASSO, which is also known under the name LSLASSO, performs better than the LASSO in terms of rate of convergence. Motivated by this result we propose to first use the LASSO to select the active variables among a large number of control variates and then to compute the OLS estimate using only the variables selected at the previous stage. We refer to this approach as the LSLASSO. To decrease the computation time when the dimensions involved in the problem, either n or m, are large, we recommend to use sub-sampling of size N smaller than n when conducting the first step.

The *active set* associated to the coefficient $\beta \in \mathbb{R}^m$ is $\text{supp}(\beta) = \{j = 1,\ldots,m : \beta_j \neq 0\}$. Let $\hat{S}_N = \text{supp}(\hat{\beta}_N^{\text{lasso}}(f))$ denote the active set of control variates based on the LASSO coefficient vector defined as in (5) but using only the first N random variables $X_1,\ldots,X_N$ generated. The LSLASSO estimate $\hat{\alpha}_n^{\text{lslasso}}(f)$ of $P(f)$ is then defined as the OLS estimate in (3) based on the full sample $X_1,\ldots,X_n$ but using only the control variates h_j restricted to $j \in \hat{S}_N$, that is,

$$\left(\hat{\alpha}_n^{\text{lslasso}}(f), \hat{\beta}_n^{\text{lslasso}}(f)\right) \in \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^{\hat{\ell}}}{\arg\min} \left\| f^{(n)} - \alpha\mathbb{1}_n - H_{\hat{S}_N}^{(n)}\beta \right\|_2^2$$

where $H_{\hat{S}_N}^{(n)}$ is the $n \times \hat{\ell}$ matrix $(h_j(X_i))_{i=1,\ldots,n,\, j\in\hat{S}_N}$ and $\hat{\ell}$ is the cardinality of $\hat{S}_N$.

Remark 8 **(Computation time)** The number of operations needed for the LSLASSO is of the order $ND + n\hat{\ell}^2 + \hat{\ell}^3 + nt$, combining the cost of selecting the control variates on the subsample of size N via cyclical coordinate descent as in Remark 7 and running the OLS estimate based on the selected control variates for the full sample of size n as in Remark 4.

3 Non-asymptotic bounds

To derive concentration inequalities for the errors of the estimators proposed in Sect. 2, we use the notion of sub-Gaussianity as defined for instance in (Boucheron et al. 2013, Sect. 2.3). Recall that the moment generating function of a centered Gaussian random variable with variance σ^2 is equal to $\lambda \mapsto \exp(\lambda^2\sigma^2/2)$.

Definition 1 A centered random variable Y is *sub-Gaussian* with variance factor $\tau^2 > 0$, notation $Y \in \mathscr{G}(\tau^2)$, if and only if $\log \mathbb{E}[\exp(\lambda Y)] \leq \lambda^2\tau^2/2$ for all $\lambda \in \mathbb{R}$.

If $Y \in \mathscr{G}(\tau^2)$, then necessarily $\text{Var}(Y) \leq \tau^2$ (Boucheron et al. 2013, Exercise 2.16). Chernoff's inequality provides exponential bounds on the tails of sub-Gaussian random

variables. Moreover, the sum of independent sub-Gaussian variables is again sub-Gaussian. Centered, bounded random variables taking values in an interval $[a, b]$ are sub-Gaussian with variance factor at most $(b-a)^2/4$ (Boucheron et al. 2013, Lemma 2.2).

The concentration inequalities for the various Monte Carlo methods with control variates will be largely due to the following assumption that requires the residuals to be sub-Gaussian.

Assumption 1 (Sub-Gaussian residuals) The residual function $\epsilon = f - P(f) - \beta^\star(f)^T h$ satisfies $\epsilon \in \mathscr{G}(\tau^2)$ for some $\tau > 0$, that is, $\int_{\mathscr{X}} \exp\{\lambda\epsilon(x)\}\, P(dx) \le \exp(\lambda^2\tau^2/2)$ for all $\lambda \in \mathbb{R}$.

The estimation error of the oracle estimator in (1) is just $\hat{\alpha}_n^{\mathrm{or}}(f) - P(f) = P_n(\epsilon) = n^{-1}\sum_{i=1}^n \epsilon(X_i)$. Under Assumption 1, this is a sub-Gaussian variable with variance factor τ^2/n. Chernoff's inequality (Boucheron et al. 2013, p. 25) then implies that for all $\delta \in (0,1)$ and all integer $n \ge 1$, with probability at least $1-\delta$,

$$\left|\hat{\alpha}_n^{\mathrm{or}}(f) - P(f)\right| \le \sqrt{2\log(2/\delta)}\frac{\tau}{\sqrt{n}}. \tag{6}$$

This concentration inequality provides a baseline when the best possible control variate in the space $\mathscr{F}_m$ is selected. The case $m = 0$ also covers the basic MC method: in that case, τ^2 is the variance factor of the sub-Gaussian variable $f - P(f)$ on $(\mathscr{X}, \mathscr{A}, P)$.

Assumption 2 (Bounded control variates) The control variates $h_1, \ldots, h_m \in L_2(P)$ are uniformly bounded. Put $U_h := \max_{j=1,\ldots,m} \sup_{x\in\mathscr{X}} \left|h_j(x)\right|$.

For a symmetric real matrix A, let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote its smallest and largest eigenvalues, respectively.

Assumption 3 (Linear independence of control variates) The control variates $h_1, \ldots, h_m \in L_2(P)$ are linearly independent. As a consequence, the $m \times m$ Gram matrix $G := P(hh^T)$ is positive definite and its smallest eigenvalue $\gamma := \lambda_{\min}(G)$ is positive.

Consider the ortho-normalized vector of control variates $\hbar = (\hbar_1, \ldots, \hbar_m)^\top = G^{-1/2}h$ and put

$$B = \sup_{x\in\mathscr{X}} h(x)^\top G^{-1} h(x) = \sup_{x\in\mathscr{X}} \hbar(x)^\top \hbar(x), \tag{7}$$

a finite quantity by Assumptions 2 and 3. The error OLS estimation error is subject to the following concentration bound.

Theorem 1 (Concentration inequality for OLS) *Suppose Assumptions 1, 2 and 3 hold. Then for all $\delta \in (0,1)$ and all integer n such that*

$$n \ge \max\bigl(18B\log(4m/\delta),\ 75m\log(4/\delta)\bigr)$$

we have, with probability at least $1-\delta$,

$$\left|\hat{\alpha}_n^{\mathrm{ols}}(f) - P(f)\right| \le \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + 58\sqrt{Bm\log(8m/\delta)\log(4/\delta)}\frac{\tau}{n}. \tag{8}$$

Compared to the bound (6) for the oracle estimator, the bound (8) for the OLS estimator has an additional term. This term is due to the additional learning step that is needed to estimate the optimal control variate.

Remark 9 **(On the factor B)** Defined as the supremum of the *leverage function* q_n in Portier and Segers (2019), Eq. (14), the quantity B plays an important role in our analysis as well as in other regression studies (Hsu et al. 2012; Newey 1997). Just as the OLS estimate (see Remark 3), the quantity B remains invariant by invertible linear transformation of the control variates. We have

$$m \le B \le \sup_{x\in\mathscr{X}} h^\top(x)h(x)/\gamma \le mU_h^2/\gamma.$$

Remark 10 **(On the parameters τ and γ)** The parameter τ in Assumption 1 is by definition an upper bound of the residual variance $\sigma_m^2(f)$. In many situations, its value is not too far from $\sigma_m^2(f)$. Hence, τ should capture the adequacy between the control variate space and the integrand f and should decrease with m. The full rank condition expressed in Assumption 3 is not crucial as one could work with the Moore–Penrose inverse when solving (4). More importantly, a large value of the minimal eigenvalue γ of the Gram matrix G reflects that the OLS problem is well-conditioned, enhancing numerical stability. As control functions are added, rows and columns are added to G and so γ cannot increase. For the Fourier basis in Example 1, we have $\gamma = 1$, while for the Legendre polynomials in Example 2, we have $\gamma \simeq 1/m$.

Remark 11 **(Link with OLS prediction risk analysis)** The approach taken in the proof of Theorem 1 requires to bound what is called the prediction risk, defined as $\|G^{1/2}(\hat{\beta}_n^{\mathrm{ols}}(f) - \beta^\star(f))\|_2$. With probability greater than $1-\delta$, we obtain an upper bound of order $\sqrt{B\tau^2\log(m/\delta)/n}$ on the prediction risk. This makes our approach comparable to the one of the recent study (Hsu et al. 2012) where concentration bounds for the OLS prediction risk (and ridge) with random design are established. In contrast to their bound, our bound involves the quantity B which shares the same invariant property as the OLS estimate and we don't require the noise to be sub-Gaussian *conditionally on the covariate* but just sub-Gaussian which is weaker.

Remark 12 **(Rates)** Consider an asymptotic set-up where the number of control variates m tends to infinity with the Monte Carlo sample size n. The OLS method improves upon the

basic MC method ($m = 0$), which has rate $1/\sqrt{n}$, as soon as $\tau + \tau\sqrt{mB\log(m)/n} \to 0$. To recover the same order as the one of the oracle estimator $\hat{\alpha}_n^{\mathrm{or}}(f)$, which has rate $\tau/\sqrt{n}$, one must have $mB\log(m) = O(n)$ as $n \to \infty$, that is, m must not be too large compared to n.

Remark 13 **(Leverage condition)** Theorem 1 may be seen as a non-asymptotic version of the asymptotic results provided in Portier and Segers (2019) in which the *leverage condition*, $\sup\{h(x)^T G^{-1} h(x) : x \in \mathscr{X}\} = o(n/m)$, is required to obtain a similar (asymptotic) bound (see Theorem 1 therein) as the one of Theorem 1. In the present non-asymptotic version, the *leverage condition* is expressed through mB when requiring that $18B\log(4m/\delta) \leq n$.

LASSO takes advantage of *sparse* regression models. A regression model is sparse whenever many of the coefficients of the parameter vector β are equal to zero, i.e., many of the covariates are useless to predict the output in the presence of the other covariates. The number of elements in the active set of the vector of regression coefficients $\beta^\star(f)$,

$$S^\star := \mathrm{supp}(\beta^\star(f)),$$

is denoted by $\ell^\star := |S^\star|$ and quantifies the level of sparsity associated to the regression model. To avoid trivialities, we tacitly assume that $S^\star$ is non-empty, so $\ell^\star \geq 1$. The factor $\ell^\star$ represents the level of sparsity of f with respect to the control functions and plays an important role in describing the benefits of the LASSO over the OLS. No assumption is made on $\ell^\star$, which could be any integer in $\{1, \ldots, m\}$.

We follow the approach presented in (Tibshirani et al. 2015, Sect. 11.4.1) (see also Bickel et al. (2009); van de Geer and Bühlmann (2009)), in which the analysis of the LASSO is carried out using a *restricted eigenvalue condition*. For a vector $\beta \in \mathbb{R}^m$ and for a non-empty set $S \subset \{1, \ldots, m\}$, write $\beta_S = (\beta_k)_{k\in S}$, seen as a (column) vector in $\mathbb{R}^{|S|}$. Define a collection of cones of interest. For $\alpha > 0$ and $S \subset \{1, \ldots, m\}$, we set $\overline{S} = \{1, \ldots, m\} \setminus S$ and

$$\mathscr{C}(S;\alpha) = \{u \in \mathbb{R}^m : \|u_{\overline{S}}\|_1 \leq \alpha \|u_S\|_1\}.$$

Assumption 4 **(Restricted eigenvalue condition)** There exists $\gamma^\star > 0$ such that $u^T G u \geq \gamma^\star \|u\|_2^2$ for all $u \in \mathscr{C}(S^\star; 3)$.

In practice, we do not know the active set $S^\star$, so the only way to ensure Assumption 4 is to make sure all control variates $h_1, \ldots, h_m$ are linearly independent. The practical value of the assumption is that $\gamma^\star \geq \gamma$, yielding sharper bounds below.

Recall that the ℓ_1-penalty of the LASSO is weighted by a regularization parameter $\lambda > 0$.

Theorem 2 (Concentration inequality for LASSO) *Suppose Assumptions 1, 2 and 4 hold. Introduce* $\xi = \ell^\star(U_h^2/\gamma^\star)$. *Then for all* $\delta \in (0,1)$ *and all integer n such that*

$$n \geq \max\left(8\xi^2\log(8m^2/\delta); 128\xi\log(8m/\delta)\right),$$
$$\lambda \geq 7U_h\sqrt{\log(8m/\delta)}\tau/\sqrt{n}$$

we have, with probability at least $1-\delta$,

$$\left|\hat{\alpha}_n^{\mathrm{lasso}}(f) - P(f)\right| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + 68\lambda\ell^\star\sqrt{\log(8m/\delta)}\frac{U_h/\gamma^\star}{\sqrt{n}}. \quad (9)$$

For λ *equal to the lower bound, we have on the same event*

$$\left|\hat{\alpha}_n^{\mathrm{lasso}}(f) - P(f)\right| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + 476\ell^\star\log(8m/\delta)(U_h^2/\gamma^\star)\frac{\tau}{n}. \quad (10)$$

Remark 14 **(LASSO vs OLS)** The benefits of LASSO over OLS can be observed by comparing the bounds in (8) and (10). The total number m, of control functions has been replaced by the active number $\ell^\star$ of such functions. Further, because $\Gamma_{S^\star} = \{u \in \mathbb{R}^p : \|u\|_2 = 1,\ u \in \mathscr{C}(S^\star;3)\}$ is included in the unit sphere, $\gamma^\star = \inf_{u\in\Gamma_{S^\star}} u^T G u$ in Assumption 3 is at least as large as the smallest eigenvalue of G, $\gamma = \inf_{\|u\|_2=1} u^T G u$ in Assumption 4.

The theoretical analysis of the LSLASSO estimator depends on the success of the LASSO-based model selection, i.e., the LASSO needs to correctly recover all the components of the true model. To ensure this selection step, the restricted eigenvalue condition is replaced by the two following ones.

Assumption 5 **(Linear independence of active functions)** The active control variates $h_k, k \in S^\star$, are linearly independent. As a consequence, the $\ell^\star \times \ell^\star$ Gram matrix $G_{S^\star} = P(h_{S^\star}h_{S^\star}^T)$ is positive definite and its smallest eigenvalue $\gamma^{\star\star} := \lambda_{\min}(G_{S^\star})$ is strictly positive.

Note that because $\{u \in \mathbb{R}^p : \|u\|_2 = 1,\ \forall k \notin S,\ u_k = 0\} \subset \Gamma_S$ (introduced in remark 14), we have that $\gamma^{\star\star} \geq \gamma^\star$. Finally, it is required that that the active control functions are orthogonal, in $L_2(P)$, to the inactive ones.

Assumption 6 **(Orthogonality)** We have $P(h_j h_k) = 0$ for all $j \in \{1, \ldots, m\} \setminus S^\star$ and all $k \in S^\star$.

Since we do not know $S^\star$ in practice, the way to ensure Assumption 6 is by making all control variates orthogonal: $P(h_j h_k) = 0$ for all $j, k \in \{1, \ldots, m\}$. The Gram matrices G and $G^\star$ are then diagonal. In the absence of zero control variates, Assumptions 3 and 4 are then satisfied as well, with $\gamma^{\star\star} = \min_{k\in S^\star} P(h_k^2) \geq \min_{k=1,\ldots,m} P(h_k^2) = \gamma > 0$.

Theorem 3 (Support recovery of LASSO) *Suppose Assumptions 1, 2, 5 and 6 hold. Then for all $\delta \in (0,1)$, all integer n such that*

$$n \geq 70(\ell^\star U_h^2/\gamma^{\star\star})^2 \log(10\ell^\star m/\delta),$$

and all λ such that

$$13U_h\sqrt{\log(10m/\delta)}\frac{\tau}{\sqrt{n}} \leq \lambda \leq \frac{\gamma^{\star\star}}{3\sqrt{\ell^\star}}\min_{k\in S^\star}|\beta_k^\star(f)|, \tag{11}$$

it holds that, with probability at least $1-\delta$, the LASSO based solution $\hat{\beta}_n^{\text{lasso}}(f)$ is unique and the true active set is recovered, $\text{supp}(\hat{\beta}_n^{\text{lasso}}(f)) = S^\star$.

The upper and lower bounds on λ in (11) must not contradict each other, and this effectively implies an additional lower bound on n. Define $B^\star = \sup_{x\in\mathscr{X}} h_{S^\star}^T(x)G_{S^\star}^{-1}h_{S^\star}(x)$ and note that

$$B^\star \leq \lambda_{\max}(G_{S^\star}^{-1}) \sup_{x\in\mathscr{X}} h_{S^\star}^T(x)h_{S^\star}(x) \leq \ell^\star U_h^2/\gamma^{\star\star}. \tag{12}$$

Theorem 4 (Concentration inequality for LSLASSO) *Suppose Assumptions 1, 2, 5 and 6 hold. Write $\xi^\star = \ell^\star(U_h^2/\gamma^{\star\star})$. Then for all $\delta \in (0,1)$ and all integer $N \in \{1,\dots,n\}$ such that*

$$N \geq 75\xi^{\star 2}\log(20\ell^\star m/\delta),$$

and all λ such that

$$13U_h\sqrt{\log(20m/\delta)}\frac{\tau}{\sqrt{N}} \leq \lambda \leq \frac{\gamma^{\star\star}}{3\sqrt{\ell^\star}}\min_{k\in S^\star}|\beta_k^\star(f)|,$$

we have, with probability at least $1-\delta$,

$$\left|\hat{\alpha}_n^{\text{lslasso}}(f) - P(f)\right| \leq \sqrt{2\log(16/\delta)}\frac{\tau}{\sqrt{n}} + 58\sqrt{B^\star\ell^\star\log(16\ell^\star/\delta)\log(8/\delta)}\frac{\tau}{n}. \tag{13}$$

The logic behind Theorem 4 is that, by Theorem 3, the active set $\hat{S}_N = \text{supp}(\hat{\beta}_N^{\text{lasso}}(f))$ identified by means of the subsample of size N is equal to the true active set $S^\star = \text{supp}(\beta^\star(f))$ with large probability. On the event that the two sets coincide, the LSLASSO estimator is then the same as the OLS estimator based on the active control variates only, and the error bound follows from Theorem 1. In practice, it turns out that LSLASSO works well even when the true active set is not identified perfectly. However, to show this formally remains an open problem.

The assumptions and concentration inequalities in our theorems feature explicit rather than generic constants. Although we have worked hard to keep these constants under control [see in particular the proof of Lemma 4 as well as Step 6(ii) in the proof of Theorem 3], it is likely that, at the cost of lengthier computations, sharper constants can still be found.

Remark 15 **(Bounded control variates)** In Assumption 2, the control variates were assumed to be bounded. Even if this assumption is valid for the two classic families in Examples 1 and 2 below, it might fail when control variates are produced with the Stein's method as suggested in Remark 1. The boundedness assumption is needed to keep the same variance factor τ^2 in the sub-Gaussian property of both variables $\epsilon(X_1)$ and $\epsilon(X_1)h(X_1)$; see, e.g., Step 3.2 in the proof of Theorem 1 or Equation (35) in the proof of Theorem 2. Avoiding this assumption is thus possible at the price of more specific assumptions on the sub-Gaussianity of $\epsilon(X_1)h(X_1)$. Note finally that (different) asymptotic results are valid for unbounded control variates (Portier and Segers 2019).

Remark 16 **(Overfitting)** Theorems 2 and 4 advocate the use of the LASSO in favor of the OLS in scenarios where ℓ^* is smaller than m or in the presence of collinearities in the design matrix making the parameter γ close to zero; see also Remark 14. Another notable advantage of the (LS)LASSO and more generally of penalization methods, is the ability to prevent over-fitting. This occurs when the number of control variates m is large compared to the Monte Carlo sample size n or, more generally, when the approximation space is large compared to the sample size. While the theory developed here is unable to address such phenomena, one of the objectives of the numerical experiments conducted in the next section is to empirically demonstrate the superior performance of the LASSO-based methods even in the absence of sparsity.

To illustrate the application of our results in a standard framework, we consider two classic families of control functions, the Fourier basis and the Legendre polynomials.

Example 1 **(Fourier basis)** On $\mathscr{X} = [0,1]$ equipped with the uniform distribution P, let $h_j(x)$ be equal to $\sqrt{2}\cos((j+1)\pi x)$ is j is odd and to $\sqrt{2}\sin(j\pi x)$ is j is even. The Fourier basis is orthonormal so that the Gram matrix is the identity, $G = I_m$, and $\gamma = \gamma^\star = \gamma^{\star\star} = 1$. The cosine and sine functions being bounded by 1, a uniform bound is $U_h = \sqrt{2}$, which implies $B \leq 2m$, $B^\star \leq 2\ell^\star$. Under the proper assumptions, we get from Theorems 1 and 4 that with probability at least $1-\delta$, since $58\sqrt{2} < 83$,

$$\left|\hat{\alpha}_n^{\text{ols}}(f) - P(f)\right| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + 83m\sqrt{\log(8m/\delta)\log(4/\delta)}\frac{\tau}{n}$$

and

$$\left|\hat{\alpha}_n^{\text{lslasso}}(f) - P(f)\right| \leq \sqrt{2\log(16/\delta)}\frac{\tau}{\sqrt{n}} + 83\ell^\star\sqrt{\log(16\ell^\star/\delta)\log(8/\delta)}\frac{\tau}{n}.$$

Example 2 **(Legendre polynomials)** Suppose that $h_j = L_j$ is the Legendre polynomial of degree $j \in \{1, \ldots, m\}$. The Legendre polynomials are orthogonal on $\mathscr{X} = [-1, 1]$ with respect to the uniform distribution P and satisfy $|L_j(x)| \leq 1$ for $x \in [-1, 1]$ with $L_j(1) = 1$ and

$$\int_{-1}^{1} L_i(x)L_j(x)\,\mathrm{d}x = \frac{2}{2j+1}\delta_{ij}.$$

The Gram matrix $G = P(hh^T)$ is diagonal with entries $1/(2j + 1)$, so the minimum eigenvalue is $\gamma = 1/(2m + 1)$ and a uniform bound is $U_h = 1$. Consequently, $B \leq 2m + 1$.

Similarly, considering only active control variates, we have $U_h^\star = 1$, while the smallest eigenvalue, $\gamma^{\star\star}$, of $G_{S^\star}$ satisfies

$$1/(2m + 1) \leq \gamma^{\star\star} \leq 1/(2\ell^\star + 1)$$

Under suitable assumptions, we get from Theorems 1 and 4 that with probability at least $1 - \delta$,

$$\left|\hat{\alpha}_n^{\text{ols}}(f) - P(f)\right| \leq \sqrt{2\log(8/\delta)}\frac{\tau}{\sqrt{n}} + 58\sqrt{(2m+1)m\log(8m/\delta)\log(4/\delta)}\frac{\tau}{n},$$

$$\left|\hat{\alpha}_n^{\text{lslasso}}(f) - P(f)\right| \leq \sqrt{2\log(16/\delta)}\frac{\tau}{\sqrt{n}} + 58\sqrt{(2\ell^\star+1)\ell^\star\log(16\ell^\star/\delta)\log(8/\delta)}\frac{\tau}{n}.$$

Compared to the Fourier basis, the improvement of LSLASSO over the OLS estimator is not only related to the number of active varables ℓ^* compared to m but also to the place of the active variables within the set of Legendre polynomials.

4 Numerical illustration

To compare the finite-sample performance of the various control variate methods, we consider synthetic data examples involving the standard integration problem over the unit cube $[0, 1]^d$. The goal is to compute $\int_{[0,1]^d} f(x)\,\mathrm{d}x$. We shall consider various dimensions $d \geq 1$, different integrands $f : [0, 1]^d \to \mathbb{R}$, and several choices for the Monte Carlo sample size, n, and the number of control variates, m. We shall focus on difficult situations where d is relatively large compared to n. In Sect. 5, we turn to real data examples in the context of Bayesian inference. For the sake of reproducibility, the data and Python code are available online[1].

Methods in competition. We consider all the methods presented in Sect. 2 with two different strategies regarding the sub-sample size used to compute the active set in LSLASSO. The methods in competition are OLS, LASSO, LSLASSO (sub-sample size $N = n$) and LSLASSOX (sub-sample size $N = \lfloor 15\sqrt{n} \rfloor$). The latter choice accelerates the computation in a substantial manner without deteriorating too much the support recovery property of the LASSO. For synthetic data, because the integration domain is the unit cube $[0, 1]^d$, Quasi-Monte Carlo (QMC) methods (Caflisch 1998) are suitable for comparison. We run such methods in the experiments with two classical low-discrepancy sets of particles, namely Halton and Sobol sequences.

On the choice of λ. In the LASSO-step of LSLASSO(X), the choice of the regularization parameter λ is essential since it controls the number of active variables. It is common to tune this parameter using K-fold cross-validation at the price of additional computations. This method, presented in general form in Algorithm 1, uses the prediction error of the underlying regression problem as a proxy to calibrate the control variates estimate. In Algorithm 1, the "data" X correspond to the matrix H of observed control variables and the "labels" y to the vector $f^{(n)}$ of observed function values. The method is computationally expensive, partitioning the training set in several folds and solving many regression problems for every value of λ in a given grid.

Algorithm 1 K-fold cross-validation

Require: data X, labels y, grid search λ_{grid}, n, K.
1. Divide $\{1, \ldots, n\}$ into K folds $F_1, \ldots, F_K$.
2. **For** $k = 1, \ldots, K$
3. Set folds $F_{-k} = \{F_1, \ldots, F_{k-1}, F_{k+1}, \ldots, F_K\}$.
4. **For** $\lambda \in \lambda_{\text{grid}}$
5. Compute estimate $\hat{\beta}_\lambda^{-k}$ on training set.
6. Compute test error $e_k(\lambda) = \sum_{i \in F_k}(y_i - x_i^T\hat{\beta}_\lambda^{-k})^2$.
7. **For** $\lambda \in \lambda_{\text{grid}}$
8. Compute average error $CV(\lambda) = \frac{1}{n}\sum_{k=1}^{K} e_k(\lambda)$.
9. **Return** $\hat{\beta}_{\lambda^\star}$ with $\lambda^\star \in \arg\min_{\lambda \in \lambda_{\text{grid}}} CV(\lambda)$.

To accelerate the computations, we suggest a new method based on a dichotomic search. Motivated by Eq. (8) and Remark 12, the value of λ is tuned such that the number of selected control variates is of the order $\sqrt{n}$, which is the order obtained for m when equating the two terms in (8) with $B = m$. Specifically, we enforce the number of activated control functions to lie in the range $[c_1\sqrt{n}, c_2\sqrt{n}]$ for

[1] https://github.com/RemiLELUC/ControlVariateSelection.git

constants $0 < c_1 < c_2$ to be chosen (see below). This choice offers two advantages. On the one hand, the upper bound $c_2\sqrt{n}$ ensures that the number of selected control variates is relatively small compared to the sample size n, promoting stability and fast computation in the final OLS step. On the other hand, the lower bound $c_1\sqrt{n}$ reduces the risk of excluding relevant control variates.

The full procedure for the selection of the regularization parameter using a dichotomic search is described below in Algorithm 2. In all experiments, we set $c_1 = 3$ and $c_2 = 12$. We initialize $\lambda = \lambda_\infty$ to be the smallest value of λ for which $\hat{\beta}^{\text{lasso}} = 0$, that is, $\lambda_\infty = \max_{k=1,\dots,m} |H_{c,k}^{(N)T} f_c^{(N)}|/N$, where $H_{c,k}^{(N)}$ stands for the k-th column of $H_c^{(N)}$, which is the same as the matrix H_c but then based on the first N Monte Carlo draws (Tibshirani et al. 2015, Exercise 2.1). Next, we decrease the value of λ, e.g., by dividing it by two, such as to incorporate more and more control variates. If too many control functions are selected, i.e., more than $c_2\sqrt{n}$, we increase the value of λ again, e.g., by multiplying it by two, to finally reach the desired range for the number of active variables. In the end, this procedure ensures a straightforward computation of the LSLASSO(X) because the size of the associated linear system remains reasonable. Contrary to K-fold cross-validation, it is not necessary to split the data into multiple folds, leading to a reduced computation time.

Algorithm 2 Dichotomic Search

Require: $f_c^{(n)}$, H_c, n, $N \le n$, (c_1, c_2).

1. Initialize $\lambda = \lambda_\infty$ and $\hat{\ell} = 0$.
2. **While** $\hat{\ell} \notin [c_1\sqrt{n}, c_2\sqrt{n}]$
3. $\quad \hat{\beta}_N^\lambda(f) \in \underset{\beta\in\mathbb{R}^m}{\arg\min} \frac{1}{2N} \|f_c^{(N)} - H_c^{(N)}\beta\|_2^2 + \lambda \|\beta\|_1$.
4. $\quad \hat{S}_N = \text{supp}(\hat{\beta}_N^\lambda(f))$ and $\hat{\ell} = \left|\hat{S}_N\right|$.
5. $\quad$ **if** $\hat{\ell} < c_1\sqrt{n}$ **then** decrease λ.
6. $\quad$ **if** $\hat{\ell} > c_2\sqrt{n}$ **then** increase λ.
7. **Return** $\hat{\beta}_N^\lambda(f)$.

The pseudo-code of the corresponding LSLASSO(X) method is provided in Algorithm 3. The regression coefficients $\hat{\beta}_n^{\text{ols}}$ and $\hat{\beta}_n^{\text{lasso}}$ for OLS and LASSO are computed using the Scikit-Learn library (Pedregosa et al. 2011), employing coordinate descent to solve the LASSO problem.

Integrands. We consider several integrands f on $[0, 1]^d$:

$$\varphi(x_1, \dots, x_d) = 1 + \sin\left(\pi\left(\frac{2}{d}\sum_{i=1}^{d} x_i - 1\right)\right), \tag{14}$$

Algorithm 3 Least-Squares Lasso Monte-Carlo (LSLASSO)

Require: $f : \mathscr{X} \to \mathbb{R}$, $h_j : \mathscr{X} \to \mathbb{R}$, $1 \le j \le m$, P, n, $N \le n$.

1. Generate $(X_i)_{i=1,\dots,n}$ independently according to P.
2. $f^{(n)} = (f(X_1), \dots, f(X_n))$ and $H = \left(h_j(X_i)\right)_{i=1,\dots,n}^{j=1,\dots,m}$.
3. $f_c^{(n)} = f^{(n)} - \mathbb{1}_n(\mathbb{1}_n^T f^{(n)})/n$ and $H_c = H - \mathbb{1}_n(\mathbb{1}_n^T H)/n$.
4. Solve $\hat{\beta}_N^\lambda(f)$ by cross-validation or dichotomic search.
5. $\hat{S}_N = \text{supp}(\hat{\beta}_N^\lambda(f))$ and $\hat{\ell} = \left|\hat{S}_N\right|$.
6. Slice $n \times \hat{\ell}$ matrix $H_{c,\hat{S}_N}^{(n)} = (H_{c\,ij}^{(n)})_{i=1,\dots,n,\ j\in\hat{S}_N}$
7. $\hat{\beta}^{\text{lslasso}}(f) \in \underset{\beta\in\mathbb{R}^m}{\arg\min} \|f_c^{(n)} - H_{c,\hat{S}_N}^{(n)}\beta\|_2^2$.
8. MC estimate $\hat{\alpha}_{n,N}^{\text{lslasso}}(f) = P_n[f - \hat{\beta}^{\text{lslasso}}(f)^T h]$.

and for all $j = 1, \dots, d$,

$$f_j(x_1, \dots, x_d) = \prod_{i=1}^{j} (2/\pi)^{1/2} x_i^{-1} \mathrm{e}^{-\log(x_i)^2/2}, \tag{15}$$

$$g_j(x_1, \dots, x_d) = \prod_{i=1}^{j} \frac{\log(2)}{2^{x_i - 1}} = \log(2)^j 2^{\sum_{i=1}^{j}(1-x_i)}. \tag{16}$$

All these functions integrate to 1 on $[0, 1]^d$. The functions f_j and g_j are built using tensor products of log-normal and exponential density functions, respectively, and depend on the first j coordinates only. This construction ensures that for small j, the integrands f_j and g_j lend themselves to Monte Carlo integration based on selected control variates. In contrast, the functions φ, f_d and g_d represent more difficult situations where all the coordinates are involved and the symmetry of their role makes it harder to select some meaningful control functions. None of the integrands belongs to the linear span of the control variates constructed in the next paragraph.

Control variates. Multidimensional control functions with respect to the uniform distribution over $[0, 1]^d$ are easy to construct based on univariate ones. Let $(h_1, \dots, h_k)$ be a vector of one-dimensional control functions, i.e., $\int_0^1 h_j(x)\,\mathrm{d}x = 0$ for each $j = 1, \dots, k$. Let $h_0 = 1$ denote the constant function equal to one. Without further information on the integrand, the usual way to construct multivariate controls is by forming tensor products of the form

$$h_\ell(x_1, \dots, x_d) = \prod_{j=1}^{d} h_{\ell_j}(x_j)$$

for a multi-index $\ell = (\ell_1, \dots, \ell_d)$ in $\{0, \dots, k\}^d \backslash \{(0, \dots, 0)\}$, yielding a total number of $(k + 1)^d - 1$ control functions.

Table 1 Number of control variates m by degree threshold deg in dimension d constructed out of tensor products of k univariate polynomials

d	k	Degree threshold (deg)				
		1	3	5	10	12
3	12	3	19	55	285	454
5	10	5	55	251	3001	6157
8	3	8	164	1214	20993	36813

The sub-sample sizes N along with the bounds $c_1\sqrt{n}$ and $c_2\sqrt{n}$ are given in Table 2.

Table 2 Sample sizes n and sub-sample sizes N together with the range $[c_1\sqrt{n}, c_2\sqrt{n}]$ corresponding to the imposed number of selected control variates in LSLASSO

n	N	$\lfloor 3\sqrt{n} \rfloor$	$\lfloor 12\sqrt{n} \rfloor$
2000	700	134	536
5000	1000	212	848
10000	2000	300	1200

A drawback of such a construction is that the number of control functions grows quickly with k. Alternative approaches yielding smaller control spaces consist of imposing $\ell_j = 0$ for all but a small number (one or two, say) of coordinates $j = 1, \ldots, d$ or simply picking at random a desired number, say m, of indices $\ell = (\ell_1, \ldots, \ell_d)$.

In this study, the set of control variates at our disposal is constructed as follows. We consider different settings of dimension d with k univariate control functions in each dimension. For $j \in \{1, \ldots, k\}$, let $h_j(x) = L_j(2x - 1)$ for $x \in [0, 1]$, with L_j the univariate Legendre polynomial (Legendre function of the first kind) of degree j; see Example 2. We have $\int_0^1 h_j(x)\,\mathrm{d}x = 0$ for all $j = 1, \ldots, m$. Because the Legendre polynomials are orthogonal, they provide some numerical stability when inverting the Gram matrix. The multivariate control functions are sorted in ascending order according to the total degree $\sum_{j=1}^{d} \ell_j \in \{1, \ldots, , kd\}$ of the polynomial. In the experiments, the number of control functions m is increased by progressively including all polynomials whose total degree is lower than or equal to a fixed threshold deg.

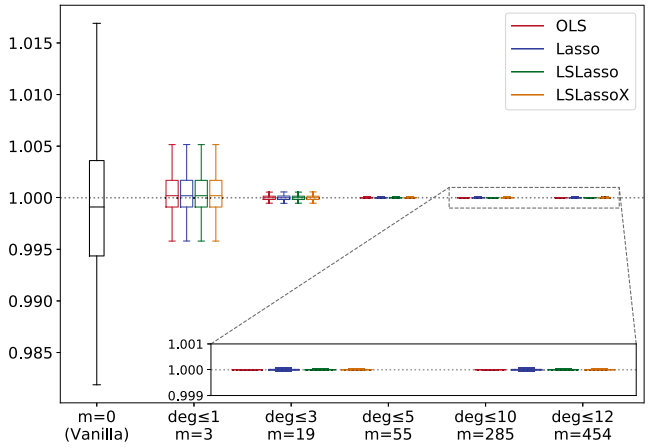
Settings. For the triple (d, k, n) we consider $d \in \{3, 5, 8\}$, $k \in \{12, 10, 3\}$ and $n \in \{2000,\ 5000,\ 10000\}$. For each choice of (d, k), the number of control variates m with a total degree lower than or equal to a fixed threshold deg are given in Table 1. The case $d = 8$ represents a difficult situation as the number of points n is relatively small compared to the dimension. For instance, a grid made of only four points in each direction would already comprise 65 536 points.

Results. The different Monte Carlo estimates are compared on the basis of their mean squared error (MSE). Figure 1 presents the boxplots obtained over 100 replications of the values returned by each of the methods. In Tables 3, 4, 5, 6, we provide the ratio $MSE(\text{vanilla})/MSE(\cdot)$, the MSE of the vanilla Monte Carlo estimate divided by the MSE for the current method, as a measure of statistical efficiency of the method relative to naive Monte Carlo integration. The four tables correspond to the four panels (a)–(d) in Fig. 1. For a given number of control variates m, the most efficient method is indicated in bold. For the Lasso-based methods, the results for the λ selection based on cross-validation (Algorithm 1) and dichotomic search (Algorithm 2) did not differ much; for the sake of brevity, the figures and the tables report the results associated to the dichotomic search.

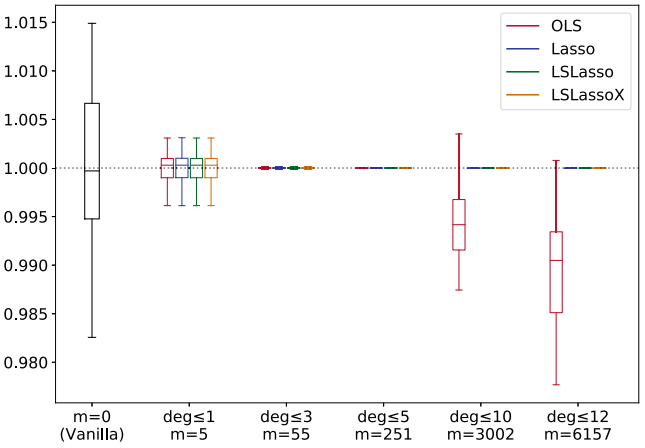
Figure 1a and b highlight the success or failure of the OLS estimator depending on the size of m compared to n. In Figure 1c and d, we consider larger values of m and only compare the Lasso-based methods as it takes too much time to solve the OLS. In all our experiments, the LSLASSOX is the clear winner as it has the highest accuracy in almost all configurations. Moreover, the LSLASSOX can be computed much faster than the LSLASSO: in our implementation, preselecting the control variates based on a smaller subsample led to a reduction op the computation time by a factor between three and twenty.

In Fig. 1a, boxplots of the values returned by each of the methods are provided for φ in (14) when $d = 3$ and $n = 10000$. In this situation, where m is small compared to n, the OLS performs very well and the LSLASSO procedure selects almost all control variates so it performs as well as OLS. In Fig. 1b, boxplots of the values returned by each of the methods are provided for g_3 in (16) when $d = 5$, $n = 2000$, and $N = 700$. In this case, the OLS estimator starts to break down as soon as the number, m, of control variates is of the same order as n. It is then necessary to perform some control variate selection, which is succesfully carried out by the LASSO and LSLASSO. Both of these estimators give the best results. Although the number of sample points used in the selection step of LSLASSOX has been reduced compared to the LSLASSO, the stability of the active set is barely affected. Accordingly, the error distributions for LSLASSO and LSLASSOX are quite similar.

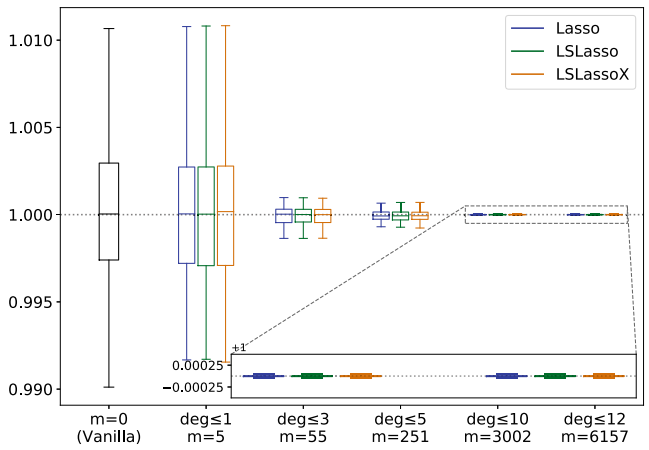
Figure 1c and d reveal the benefits of selecting appropriate control variates before applying the OLS estimator. Figure 1c covers the function f_1 in (15) when $d = 5$ and $n = 5000$, while Fig. 1d deals with the function g_4 in (16) when $d = 8$ and $n = 2000$. In the latter case, the number of control variates, $m = 36813$, is huge compared to the sample size $n = 2000$. However, the Lasso-based methods perform remarkably well in those settings. More precisely, in dimension $d = 5$ with the function f_1, the mean square error of the naive Monte Carlo estimator is of the order 10^{-5} whereas the



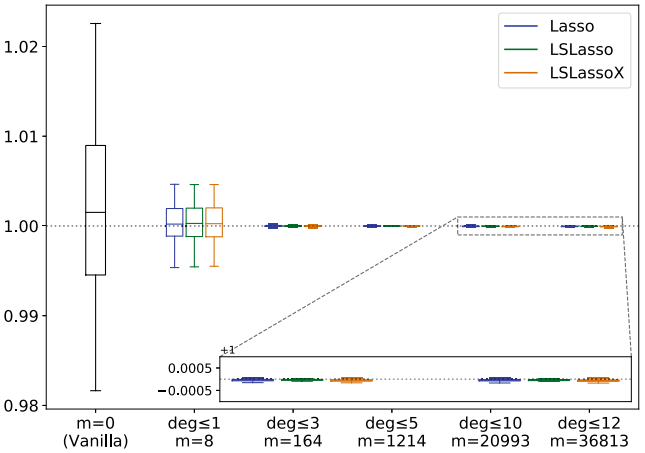
(a) Boxplots for φ in (14) with $d = 3$, $n = 10000$, $N = 2000$.



(b) Boxplots for g_3 in (16) with $d = 5$, $n = 2000$, $N = 700$.



(c) Boxplots for f_1 in (15) with $d = 5$, $n = 5000$, $N = 1000$.



(d) Boxplots for g_4 in (16) with $d = 8$, $n = 2000$, $N = 700$.

Fig. 1 Boxplots (based on 100 runs) of the values returned by each of the methods for functions φ, g_3, f_1, g_4 in (14)–(16)

Table 3 Statistical efficiency for φ; see also Fig. 1a

$m =$	3	19	55	285	454
OLS	8.42e00	8.56e02	**2.12e05**	2.49e11	**5.27e14**
LASSO	8.42e00	8.53e02	6.72e04	7.71e04	7.71e04
LSL	8.42e00	**8.58e02**	2.10e05	6.26e05	1.37e06
LSLX	8.42e00	8.51e02	2.09e05	**2.49e11**	2.91e05
QMC	Halton: 8.76e01		Sobol: 3.29e02		

Table 4 Statistical efficiency for g_3; see also Fig. 1b

$m =$	5	55	251	3002	6157
OLS	2.45e01	5.75e04	7.48e08	1.42e00	4.94e-1
LASSO	2.45e01	5.75e04	4.19e06	4.83e05	4.31e05
LSL	2.45e01	5.75e04	**7.79e08**	**4.83e06**	**4.54e06**
LSLX	2.45e01	5.75e04	1.87e08	1.71e06	5.54e05
QMC	Halton: 3.75e00		Sobol: 1.57e01		

Table 5 Statistical efficiency for f_1; see also Figure 1c

$m =$	5	55	251	3002	6157
LASSO	1.11e00	**6.60e01**	**1.79e02**	8.17e04	8.56e04
LSL	1.11e00	6.59e01	1.76e02	6.77e04	6.83e04
LSLX	1.11e00	6.59e01	1.78e02	**8.97e04**	**9.24e04**
QMC	Halton: 4.60e00		Sobol: 7.21e01		

Table 6 Statistical efficiency for g_4; see also Fig. 1d

$m =$	8	164	1214	20993	36813
LASSO	1.98e01	1.52e04	7.94e05	7.94e04	6.05e04
LSL	1.97e01	1.53e04	1.32e06	**1.49e05**	**1.28e05**
LSLX	1.98e01	**1.54e04**	**1.38e06**	1.98e04	1.55e04
QMC	Halton: 3.80e00		Sobol: 2.60e01		

one of the LSLASSOX is of the order 10^{-10}. Similarly, in dimension $d = 8$ with the function g_4, the mean square error goes down from 10^{-4} to 10^{-8}. Table 6 highlights the benefits of the LSLASSO over the LASSO in difficult situations.

In the recent study (South et al. 2018), the authors investigate the use of *regularization* in computing control variates estimates. They focus on the LASSO and ridge regression and they show, based on several examples, that the LASSO generally outperforms the ridge. In the applications they consider, they found that polynomials with relatively small degrees in each direction (k equal to 2 and 3) give the best performance. The examples considered here show a similar pattern as the results do not generally improve beyond degree $k = 3$.

5 Bayesian inference

In this section, we compare the different Monte Carlo estimates on Bayesian inference examples. Given some observed data x, the goal is to infer the parameter θ of a statistical model. We have some information through the prior distribution $\pi(\theta)$ and observe the model likelihood $\ell(x|\theta)$. Bayes' rule gives the posterior distribution as

$$p(\theta|x) = \frac{\ell(x|\theta)\pi(\theta)}{\int_\Theta \ell(x|\theta)\pi(\theta)\mathrm{d}\theta}.$$

The normalizing constant in the denominator is called evidence and is of interest for Bayesian model selection:

$$Z = \int_\Theta \ell(x|\theta)\pi(\theta)\mathrm{d}\theta.$$

Typically, this integral is analytically intractable. It is also difficult to compute numerically if the dimension d of the parameter space Θ is large.

We consider the same datasets as in South et al. (2018): the European dipper capture-recapture data from Marzolin (1988) in Sect. 5.1 and the sonar data from Gorman and Sejnowski (1988) in Sect. 5.2. The dimensions of the integration domains are $d = 12$ and $d = 61$, respectively.

As in Sect. 4, we consider multivariate control functions based on univariate orthogonal polynomials by forming tensor products of the form $h_\ell(x_1, \ldots, x_d) = \prod_{j=1}^d h_{\ell_j}(x_j)$, for a multi-index $\ell = (\ell_1, \ldots, \ell_d)$ in $\{0, \ldots, k\}^d \setminus \{(0, \ldots, 0)\}$. In both examples, the dimension d is so large that considering all tensor products is infeasible. Instead, we focus on combinations where ℓ_j equals 0 for all but one or two coordinates, leading to a total number of $m = kd$ and $m = kd + k^2 d(d-1)/2$ control variates, respectively.

The different Monte Carlo estimates are compared on the basis of their mean squared errors (MSE). In contrast to Sect. 4, the true value of the integral is unknown. An estimate of this value, referred to as the gold standard $Z^\star$, is obtained by naive Monte Carlo with sample size $n = 10^8$. The variance of this estimate, computed on 20 independent runs, is smaller than the variance of all the other considered methods. The different boxplots of Fig. 2 show the results obtained over 100 independent runs of $\hat{Z}/Z^\star$ where $\hat{Z}$ is the estimate of the evidence. Tables 8, 9, 10, 11 provide numerical values for the statistical efficiency $\widehat{MSE}(\text{vanilla})/\widehat{MSE}(\cdot)$. We consider various settings and the parameter configuration is $n \in \{2\,000;\ 5\,000\}$ for the Monte Carlo sample size with $N \in \{700;\ 1\,000\}$ for the Monte Carlo subsample size for the LSLASSO. The regularization parameter λ is chosen via dichotomic search (Algorithm 2).

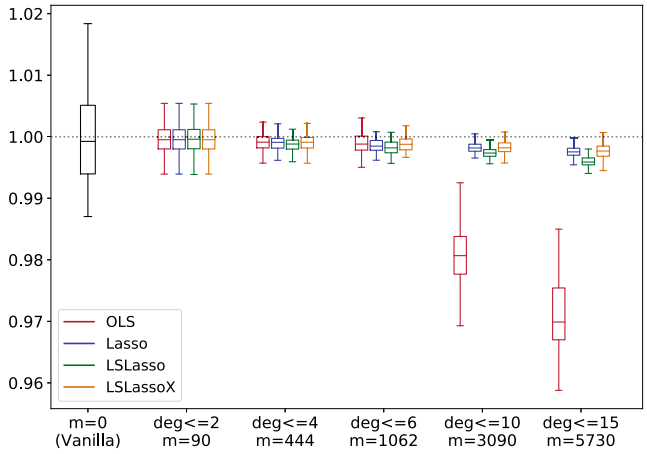
Table 7 European dipper capture-recapture data (Marzolin 1988). The counts in the triangle refer to the number of birds released in a given year and recaptured for the first time in a later year

Release year	Birds released	Year of recapture: 1981 + ⋯					
		1	2	3	4	5	6
1981	22	11	2	0	0	0	0
1982	60		24	1	0	0	0
1983	78			34	2	0	0
1984	80				45	1	2
1985	88					51	0
1986	98						52

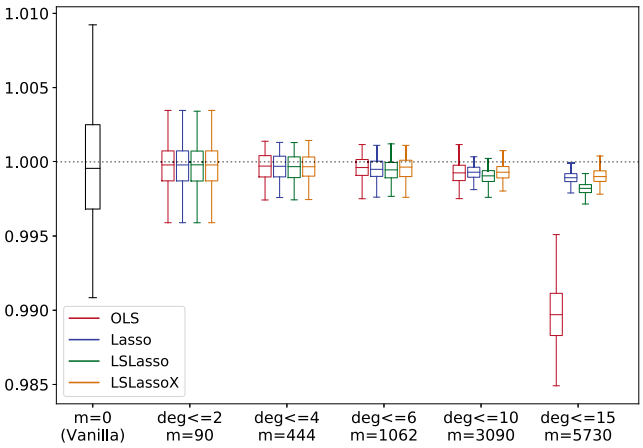
5.1 European dipper capture-recapture data

The data-set given in Table 7 was collected by Marzolin (1988) and describes the annual capture and recapture counts of the bird species *Cinclus cinclus*, also known as the European dipper, in eastern France from 1981 to 1987. We observe count data $x_{i,j}$ with $i \in \{1, \ldots, I\}$ and $j \in \{i+1, \ldots, J\}$, where $x_{i,j}$ denotes the number of birds released in year i and subsequently recaptured for the first time in year j. In the example, we have $I = 6$ and $J = 7$, where 1981 corresponds to year $i = 1$. Also observed is R_i, the number of marked birds released into the population in year i.

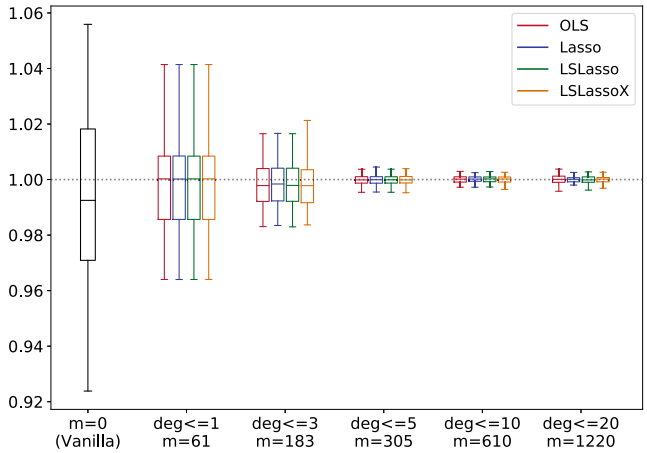
Following (Brooks et al. 2000; Nott et al. 2018; South et al. 2018), we consider a Bayesian approach for the Cormack–Jolly–Seber model (Lebreton et al. 1992). The model parameters are ϕ_i, a bird's survival probability from year i to $(i+1)$ for $i \in \{1, \ldots, I\}$, together with p_j, the probability of a bird being recaptured in year $j \in \{2, \ldots, J\}$. Let $\nu_{i,j}$ denote the probability that a bird captured and released in year i gets recaptured for the first time in year j. Since the bird must survive from year i to year j, not be recaptured in years $i+1$ to $j-1$ and then finally be recaptured in year j, the probability is modelled as



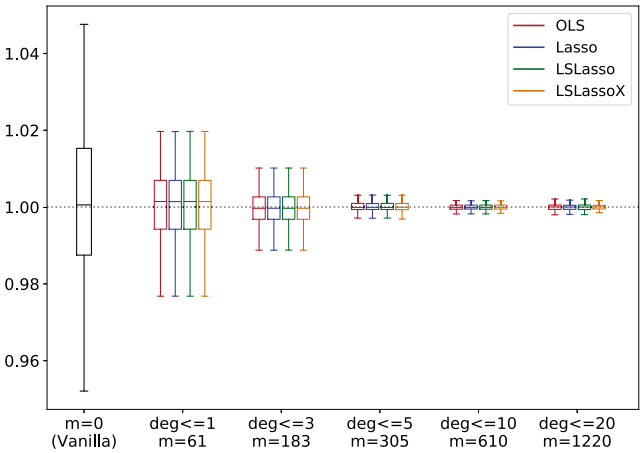
(a) Boxplots for Capture dataset ($n = 2000,\ N = 700$)



(b) Boxplots for Capture dataset ($n = 5000,\ N = 1000$)



(c) Boxplots for Sonar dataset ($n = 2000,\ N = 700$)



(d) Boxplots for Sonar dataset ($n = 5000,\ N = 1000$)

Fig. 2 Boxplots (based on 100 runs) of $\hat{Z}/Z^{\star}$ returned by each of the methods for Capture-Recapture and Sonar examples

Table 8 Capture data: statistical efficiency ($n = 2000$)

$m =$	90	444	1 062	3 090	5 730
OLS	9.33	**20.7**	14.7	0.14	0.06
LASSO	9.34	20.3	16.7	**14.4**	**8.57**
LSL	9.33	20.4	12.8	8.43	4.60
LSLX	9.33	19.4	**19.8**	12.9	7.86

Table 9 Capture data: statistical efficiency ($n = 5000$)

$m =$	90	444	1 062	3 090	5 730
OLS	7.67	18.1	22.1	15.2	0.15
LASSO	7.67	**18.4**	**22.3**	**22.8**	12.8
LSL	7.67	18.0	21.3	13.3	5.24
LSLX	7.67	17.8	21.4	21.6	**13.2**

Table 10 Sonar data: statistical efficiency ($n = 2000$)

$m =$	61	183	305	610	1220
OLS	3.39	13.3	246	548	330
LASSO	3.39	13.6	**250**	**673**	680
LSL	3.39	13.3	246	564	499
LSLX	3.39	**13.9**	244	558	**680**

Table 11 Sonar data: statistical efficiency ($n = 5000$)

$m =$	61	183	305	610	1220
OLS	4.48	17.0	235	801	601
LASSO	4.49	17.0	**240**	821	721
LSL	4.48	17.0	235	804	629
LSLX	4.48	17.0	241	**833**	**734**

$$\nu_{i,j} = \phi_i p_j \prod_{k=i+1}^{j-1} [\phi_k(1-p_k)].$$

The number of birds released at year i that are never recaptured at all is equal to $r_i = R_i - \sum_{j=i+1}^{J} x_{i,j}$ while the probability that a bird released in year i is never recaptured is $\chi_i = 1 - \sum_{j=i+1}^{J} \nu_{i,j}$. The resulting likelihood is equal to

$$\ell(x|\theta) = \prod_{i=1}^{I} \left\{ \chi_i^{r_i} \prod_{j=i+1}^{J} \nu_{i,j}^{x_{i,j}} \right\},$$

where $\theta = (\phi_1, \ldots, \phi_6, p_2, \ldots, p_7) \in [0,1]^{12}$. The uniform distribution is chosen as prior and we use tensor products of Legendre polynomials with $k = 10$ (Example 2) as controls.

The results for the various integration methos are reported in the same way as in Sect. 4. The boxplots and statistical efficiencies are given in Fig. 2a and b and Tables 8 and 9 respectively. Similarly to the synthetic data, Fig. 2a and b reveal the success or failure of the OLS on the capture-recapture data when the number of control variates m is larger than the Monte Carlo sample size n. The variance goes down as m increases. Tables 8 and 9 show that for $n = 2\,000$, the OLS estimate gives the best performance whereas for $n = 5\,000$, the LASSO-based methods profit from the large number of available control variates. In this case, the LASSO is most efficient while the LSLASSOX performs similarly but at a reduced computing time.

5.2 Sonar data

The data were collected by Gorman and Sejnowski (1988) and are available from the UCI Machine Learning Repository (Asuncion and Newman 2007). The data matrix X represents 208 sonar signals, each one composed of 60 attributes within the binary classification framework. A column of 1's is added to the matrix X to represent the intercept so that $X \in \mathbb{R}^{208\times 61}$. The goal is to assess whether the sonar signal bounces off a metal cylinder (label $y = 1$) or a roughly cylindrical rock (label $y = -1$). The different covariates represent the energy within particular frequency bands, integrated over a certain period of time. Using the encoding $y \in \{-1, +1\}$ and following a logistic regression model, the resulting log-likelihood is

$$\log \ell(X, y|\theta) = -\sum_{i=1}^{208} \log \left\{ 1 + \exp\left(-y_i \sum_{j=1}^{61} X_{ij}\theta_j \right) \right\},$$

where the model coefficient $\theta \in [-1,1]^{61}$ has a uniform prior distribution. We use the family of Legendre polynomials as control functions with $k = 20$. The boxplots and statistical efficiencies are presented in Figures 2c and d and Tables 10 and 11, respectively. Once again, the Lasso-based methods, with their selection strategy, are able to benefit from a larger control variates space. The winner of this competition is LSLASSOX as it offers the best performance combined with a smaller computation time compared to the LSLASSO.

6 Conclusion and perspective

The use of high-dimensional control variates with the help of a LASSO-type procedure has been shown to be efficient in order to reduce the variance of the basic Monte Carlo estimate. The method, called LSLASSO(X), that first selects appropriate control variates by the LASSO, possibly on a smaller subsample, and then estimates the control variate coefficients by least squares performs excellently considering the modest computing time required. Several avenues for further research are now discussed.

The construction of control variates by a change of measure (Remark 1) presupposes some knowledge on the underlying integration measure in order to choose an appropriate sampling distribution. For instance, if the support of the sampling measure does not cover the whole integration domain then the method will certainly fail. *Adaptive importance sampling* (see, e.g., Owen and Zhou (2000); Portier and Delyon (2018)) offers a possible solution, involving online estimates of the appropriate sampling policy and the optimal linear combination of control variates.

Assumption 1 on the sub-Gaussianity of the residuals is key to obtain concentration inequalities. For certain applications, it might by too restrictive, however. In the absence of such an assumption or more generally of suitable bounds on the tails of the residual distribution, other types of results such as almost sure convergence rates might still be pursued.

In the random design setting, the estimators of coefficient vector $\beta^\star(f)$ are all biased, even the OLS estimator. The bias may be removed by sample splitting (Avramidis and Wilson 1993), but at the cost of an increased variance, especially if the number of control variates is large. For the Lasso-based methods, debiasing methods are studied in Javanmard and Montanari (2018) and the references therein. The merits of these techniques for Monte Carlo control variate methods remain to be investigated.

We have presented different control variate methods from the point of view of estimation only. Equally important questions are that of model evaluation and Monte Carlo sample size calculation, assessing the accuracy of the estimate. Several ways can be imagined such as sample splitting (e.g., cross-validation) and plug-in estimation of the residual variance $\sigma^2(f)$, using for instance the estimated residuals.

Acknowledgements The authors are grateful to the Associate Editor and two anonymous Reviewers for their many valuable comments and interesting suggestions. Further, the authors gratefully acknowledge the exceptionally careful proofreading done by Aigerim Zhuman (UCLouvain).

A Auxiliary results

Lemma 1 (Sub-Gaussian) *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables in $(\mathscr{X}, \mathscr{A})$ with distribution P. Let $\varphi_1, \ldots, \varphi_p$ be real-valued functions on $\mathscr{X}$ such that $P(\varphi_k) = 0$ and $\varphi_k \in \mathscr{G}(\tau^2)$ for all $k = 1, \ldots, p$. Then for all $\delta > 0$, we have with probability at least $1 - \delta$,*

$$\max_{k=1,\ldots,p} \left| \sum_{i=1}^{n} \varphi_k(X_i) \right| \leq \sqrt{2n\tau^2 \log(2p/\delta)}.$$

Proof For each $k = 1, \ldots, p$, the centered random variable $\sum_{i=1}^{n} \varphi_k(X_i)$ is sub-Gaussian with variance factor $n\tau^2$. By the union bound and by Chernoff's inequality, we have, for each $t > 0$,

$$\mathbb{P}\left(\max_{k=1,\ldots,p} \left| \sum_{i=1}^{n} \varphi_k(X_i) \right| > t \right) \leq \sum_{k=1}^{p} \mathbb{P}\left(\left| \sum_{i=1}^{n} \varphi_k(X_i) \right| > t \right) \leq 2p \exp\left(\frac{-t^2}{2n\tau^2} \right).$$

Set $t = \sqrt{2n\tau^2 \log(2p/\delta)}$ to find the result. □

Lemma 2 (Smallest eigenvalue lower bound) *Let $X_1, \ldots, X_n$ be independent and identically distributed random variables in $(\mathscr{X}, \mathscr{A})$ with distribution P. Let $g = (g_1, \ldots, g_p)^T$ in $L_2(P)^p$ be such that the $p \times p$ Gram matrix $G = P(gg^T)$ satisfies $\lambda_{\min}(G) > 0$. Define the transformation $\tilde{g} = G^{-1/2} g$ and put $B_{\tilde{g}} := \sup_{x \in \mathscr{X}} \|\tilde{g}(x)\|_2^2$. Let $\delta, \eta \in (0, 1)$. For $\delta \in (0, 1)$, the empirical Gram matrix $\hat{G}_n = P_n(gg^T)$ satisfies, with probability at least $1 - \delta$,*

$$\lambda_{\min}(\hat{G}_n) > \left(1 - \sqrt{2B_{\tilde{g}} n^{-1} \log(p/\delta)} \right) \lambda_{\min}(G).$$

Proof Suppose that the result is true in the special case that G is the identity matrix. In case of a general Gram matrix G, we could then apply the result for the special case to the vector of functions $\tilde{g} = G^{-1/2} g$, whose Gram matrix is the identity matrix. We would get that $\lambda_{\min}(P_n(\tilde{g}\tilde{g}^T)) > 1 - \eta$ with probability at least $1 - \delta$. Since $P_n(\tilde{g}\tilde{g}^T) = G^{-1/2} \hat{G}_n G^{-1/2}$ and since $u^T G^{-1} u \leq 1/\lambda_{\min}(G)$ for every unit vector $u \in \mathbb{R}^p$, we would have

$$\begin{aligned} \lambda_{\min}\left(P_n(\tilde{g}\tilde{g}^T) \right) &= \min_{u^T u = 1} \left\{ u^T P_n(\tilde{g}\tilde{g}^T) u \right\} \\ &= \min_{u^T u = 1} \left\{ \frac{(G^{-1/2}u)^T \hat{G}_n G^{-1/2} u}{(G^{-1/2}u)^T G^{-1/2} u} u^T G^{-1} u \right\} \\ &\leq \lambda_{\min}(\hat{G}_n) / \lambda_{\min}(G). \end{aligned}$$

It would then follow that

$$\lambda_{\min}(\hat{G}_n) \geq \lambda_{\min}(P_n(\tilde{g}\tilde{g}^T))\, \lambda_{\min}(G) \geq (1 - \eta) \lambda_{\min}(G),$$

as required. Hence we only need to show the result for $G = I$, in which case $\tilde{g} = g$.

We apply the matrix Chernoff inequality in (Tropp 2015, Theorem 5.1.1) to the random matrices $n^{-1} g(X_i) g(X_i)^T$. These matrices are independent and symmetric with dimension $p \times p$. Their minimum and maximum eigenvalues are between 0 and $L = B_g/n$, with $B_g = \sup_{x \in \mathscr{X}} \lambda_{\max}(g(x)g(x)^T) = \sup_{x \in \mathscr{X}} \|g(x)\|_2^2$. Their sum is equal to $P_n(gg^T) = \hat{G}_n$, whose expectation is $G = I$ by assumption. In the notation of the cited theorem, we have $\mu_{\min} = \lambda_{\min}(G) = 1$, and thus, by Eq. (5.1.5) in that theorem, we have, for $\eta \in [0, 1)$,

$$\mathbb{P}\{\lambda_{\min}(\hat{G}_n) \leq 1 - \eta\} \leq p \left[\frac{\exp(-\eta)}{(1 - \eta)^{1 - \eta}} \right]^{n/B_g}.$$

The term in square brackets is bounded above by $\exp(-\eta^2/2)$. Indeed, we have, for $\eta \in [0, 1)$,

$$\frac{e^{-\eta}}{(1 - \eta)^{1 - \eta}} = \exp\{-\eta - (1 - \eta)\log(1 - \eta)\}$$

and

$$\begin{aligned} \eta + (1 - \eta)\log(1 - \eta) &= \eta - (1 - \eta) \int_0^{\eta} \frac{\mathrm{d}t}{1 - t} \\ &= \int_0^{\eta} \left(1 - \frac{1 - \eta}{1 - t} \right) \mathrm{d}t \\ &= \int_0^{\eta} \frac{\eta - t}{1 - t} \, \mathrm{d}t \\ &\geq \int_0^{\eta} (\eta - t) \, \mathrm{d}t = \frac{\eta^2}{2}. \end{aligned}$$

It follows that

$$\mathbb{P}\{\lambda_{\min}(\hat{G}_n) \leq 1 - \eta\} \leq p \exp\left(-\frac{\eta^2 n}{2B_g} \right).$$

Solving $p \exp\left(-\frac{\eta^2 n}{2B_g}\right) = \delta$ in η, we find that, with probability at least $1 - \delta$,

$$\lambda_{\min}(\hat{G}_n) > 1 - \sqrt{2B_g n^{-1} \log(p/\delta)}.$$

□

Lemma 3 (Upper bound of moments) *Let X be a random variable such that $\mathbb{E}(|X|^{2p}) \le 2^{p+1}p!$ for every integer $p \ge 1$. Then*

$$\forall \lambda \in \mathbb{R}, \qquad 1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(|X|^k) \le \exp(9\lambda^2/4), \tag{17}$$

in which it is implicitly understood that the series on the left-hand side converges.

Proof Let $\lambda \in \mathbb{R}$. We split the series in terms with even and odd indices k, leading to

$$\begin{aligned} &1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(|X|^k) \\ &\quad = 1 + \sum_{p=1}^{\infty} \frac{\lambda^{2p}}{(2p)!}\mathbb{E}(|X|^{2p}) + \sum_{p=1}^{\infty} \frac{\lambda^{2p+1}}{(2p+1)!}\mathbb{E}(|X|^{2p+1}). \end{aligned}$$

We will bound the series on the odd indices in terms of the series on the even indices.

Since the geometric mean of two nonnegative numbers is bounded by their arithmetic mean, we have, for all $x \ge 0$ and all $a > 0$,

$$|x| = \sqrt{\frac{1}{a} \cdot ax^2} \le \frac{1}{2}\left(\frac{1}{a} + ax^2\right).$$

Applying the previous inequality to $x = \lambda X$ and scalars $a_p > 0$ to be chosen later,

$$\begin{aligned} &\sum_{p=1}^{\infty} \frac{\lambda^{2p+1}}{(2p+1)!}\mathbb{E}(|X|^{2p+1}) \\ &\quad \le \sum_{p=1}^{\infty} \frac{\lambda^{2p}}{(2p+1)!}\mathbb{E}\left[|X|^{2p}\frac{1}{2}\left(\frac{1}{a_p} + a_p(\lambda X)^2\right)\right] \\ &\quad = \sum_{p=1}^{\infty} \frac{\lambda^{2p}}{2a_p}\frac{\mathbb{E}(|X|^{2p})}{(2p+1)!} + \sum_{p=1}^{\infty} \frac{a_p}{2}\frac{\lambda^{2p+2}}{(2p+1)!}\mathbb{E}(|X|^{2p+2}) \\ &\quad = \sum_{p=1}^{\infty} \frac{\lambda^{2p}}{2a_p}\frac{\mathbb{E}(|X|^{2p})}{(2p+1)!} + \sum_{p=2}^{\infty} \frac{a_{p-1}}{2}\frac{\lambda^{2p}}{(2p-1)!}\mathbb{E}(|X|^{2p}) \\ &\quad = \sum_{p=1}^{\infty} \left(\frac{1}{2a_p(2p+1)} + pa_{p-1}\mathbb{1}_{\{p\ge 2\}}\right)\frac{\lambda^{2p}}{(2p)!}\mathbb{E}(|X|^{2p}). \end{aligned}$$

Here, $\mathbb{1}$ denotes an indicator function. We obtain

$$\begin{aligned} &\sum_{k=2}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(|X|^k) \\ &\quad \le \sum_{p=1}^{\infty} \left(1 + \frac{1}{2a_p(2p+1)} + pa_{p-1}\mathbb{1}_{\{p\ge 2\}}\right)\frac{\lambda^{2p}}{(2p)!}\mathbb{E}(|X|^{2p}). \end{aligned}$$

Define $b_p = a_p(2p+1)$ and use the hypothesis on $\mathbb{E}(|X|^{2p})$ to see that, for any constants $b_p > 0$,

$$\begin{aligned} &1 + \sum_{k=2}^{\infty} \frac{\lambda^k}{k!}\mathbb{E}(|X|^k) \\ &\quad \le 1 + \sum_{p=1}^{\infty} \left(1 + \frac{1}{2b_p} + \frac{p}{2p-1}b_{p-1}\mathbb{1}_{\{p\ge 2\}}\right)\frac{\lambda^{2p}}{(2p)!}2^{p+1}p!. \end{aligned}$$

The objective is to find a constant $c > 0$ as small as possible and such that the right-hand side is bounded by $\exp(c\lambda^2) = 1 + \sum_{p=1}^{\infty} c^p\lambda^{2p}/p!$. Comparing coefficients, this means that we need to determine scalars $b_p > 0$ and $c > 0$ in such a way that for all $p = 1, 2, \ldots$

$$\left(1 + \frac{1}{2b_p} + \frac{p}{2p-1}b_{p-1}\mathbb{1}_{\{p\ge 2\}}\right)\frac{2^{p+1}p!}{(2p)!} \le \frac{c^p}{p!},$$

or, equivalently,

$$\left(2 + \frac{1}{b_p} + \frac{2p}{2p-1}b_{p-1}\mathbb{1}_{\{p\ge 2\}}\right)\prod_{j=1}^{p}\frac{j}{p+j} \le (c/2)^p.$$

The case $p = 1$ gives

$$2 + \frac{1}{b_1} \le c, \tag{18}$$

showing that, with this proof technique, we will always find $c > 2$. Setting $b_p \equiv b > 0$ for all integer $p \ge 1$ and $c = 2 + 1/b$, inequality (18) is automatically satisfied, so it remains to find $b > 0$ such that forall $p = 2, 3, \ldots$

$$\left(2 + \frac{1}{b} + \frac{2p}{2p-1}b\right)\prod_{j=1}^{p}\frac{j}{p+j} \le (c/2)^p \quad \text{with } c = 2 + \frac{1}{b}.$$

The left-hand side is decreasing in p whereas the right-hand side is increasing in p. It is thus sufficient to have the inequality satisfied for $p = 2$, i.e.,

$$\left(2 + \frac{1}{b} + \frac{4b}{3}\right)\frac{1}{6} \le \left(1 + \frac{1}{2b}\right)^2. \tag{19}$$

Equating both sides leads to a nonlinear equation in b that can be solved numerically, giving the root $b \approx 4.006156$. With $b = 4$, inequality (19) is satisfied, as can be checked directly ($91/72 \le 81/64$). We conclude that $c = 2 + 1/4 = 9/4$ is a valid choice. □

Note that the series in (17) starts at $k = 2$. If also $\mathbb{E}(X) = 0$, the left-hand side in (17) is an upper bound for $\mathbb{E}(\exp(\lambda X))$, and we obtain the following corollary.

Corollary 1 *Let Z be a centered random variable such that*

$$\forall p \in \mathbb{N}^\star, \quad \mathbb{E}(|Z|^{2p}) \leq 2^{p+1} p!$$

Then $\log \mathbb{E}(\exp(\lambda Z)) \leq 9\lambda^2/4$ *for all* $\lambda \in \mathbb{R}$, *i.e.,* $Z \in \mathcal{G}(9/2)$.

Lemma 4 *Let (X, Y) be a pair of uncorrelated random variables. If $X \in \mathcal{G}(v)$ and $|Y| \leq \kappa$ for some $v > 0$ and $\kappa > 0$, then $XY \in \mathcal{G}((9/2)\kappa^2 v)$.*

Proof The random variable $X/\sqrt{v}$ is sub-Gaussian with variance factor 1. As on page 25 in Boucheron et al. (2013), this implies that $\mathbb{P}(|X/\sqrt{v}| > t) \leq 2\exp(-t^2/2)$ for all $t \geq 0$ and thus $\mathbb{E}[|X/\sqrt{v}|^{2p}] \leq 2^{p+1} p!$ for all integer $p \geq 1$ (see (Boucheron et al. 2013, Theorem 2.1)).

Let $Z = XY/(\sqrt{v}\kappa)$. Since X is centered and X and Y are uncorrelated, XY is centred too, and therefore also Z. From the previous paragraph, we have $\mathbb{E}(|Z|^{2p}) \leq \mathbb{E}(|X/\sqrt{v}|^{2p}) \leq 2^{p+1} p!$ for all integer $p \geq 1$. Corollary 1 gives for all $\lambda \in \mathbb{R}$ that $\log \mathbb{E}(\exp(\lambda Z)) \leq 9\lambda^2/4$, from which

$$\log \mathbb{E}(\exp(\lambda XY)) = \log \mathbb{E}(\exp(\lambda\sqrt{v}\kappa Z)) \leq \frac{9}{4}\lambda^2 v\kappa^2.$$

□

Lemma 5 (Upper bound for norm-subGaussian random vector) *Let X be a d-dimensional random vector with zero-mean and such that $\mathbb{P}(\|X\|_2 \geq t) \leq 2\exp\left(-t^2/(2\sigma^2)\right)$ for all $t \geq 0$. Then the random matrix Y defined by*

$$Y = \begin{bmatrix} 0 & X^T \\ X & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(d+1)} \tag{20}$$

satisfies $\mathbb{E}(\exp(\theta Y)) \preceq \exp(c\theta^2\sigma^2) I$ for any $\theta \in \mathbb{R}$, with $c = 9/4$, where I denotes the identity matrix.

Proof The non-zero eigenvalues of Y are $\|X\|$ and $-\|X\|$. The non-zero eigenvalues of Y^k are thus $\|X\|^k$ and $(-\|X\|)^k$ for integer $k \geq 1$. It follows that $Y^k \preceq \|X\|^k I$ for all integer $k \geq 1$, and therefore also $\mathbb{E}(Y^k) \preceq \mathbb{E}(\|X\|^k) I$ for all integer $k \geq 1$. Furthermore, the operator norm of Y^k is bounded by $\|Y^k\| \leq \|X\|^k$.

Since $\mathbb{E}(Y) = 0$, we get, for any $\theta \in \mathbb{R}$,

$$\mathbb{E}(\exp(\theta Y)) = I + \sum_{k=2}^{\infty} \frac{\theta^k}{k!}\mathbb{E}(Y^k)$$
$$\preceq \left(1 + \sum_{k=2}^{\infty} \frac{\theta^k}{k!}\mathbb{E}(\|X\|^k)\right) I = \left(1 + \sum_{k=2}^{\infty} \frac{(\theta\sigma)^k}{k!}\mathbb{E}(\xi^k)\right) I,$$

where $\xi = \|X\|/\sigma$. The first series converges in operator norm since $\|\mathbb{E}(Y^k)\| \leq \mathbb{E}(\|Y^k\|) \leq \mathbb{E}(\|X\|^k)$.

By assumption, $\mathbb{P}(\xi > t) = \mathbb{P}(\|X\| \geq \sigma t) \leq 2e^{-t^2/2}$ for all $t \geq 0$ and thus $\mathbb{E}(|\xi|^{2p}) \leq 2^{p+1} p!$ for all integer $p \geq 1$ But then we can apply Lemma 3 with $\lambda = \theta\sigma$ and $X = \xi$, completing the proof. □

The following result is a special case of Jin et al. (2019, Corollary 7). Our contribution is to make the constant c in the cited result explicit. In passing, we correct an inaccuracy in the proof of Jin et al. (2019, Lemma 4), in which it was incorrectly claimed that the odd moments of a certain random matrix Y as in our Lemma 5 are all zero.

Lemma 6 (Hoeffding inequality for norm-subGaussian random vectors) *Let the d-dimensional random vectors $Z_1, \ldots, Z_n$ be independent, have mean zero, and satisfy*

$$\forall t \geq 0, \forall i = 1, \ldots, n, \quad \mathbb{P}(\|Z_i\|_2 \geq t) \leq 2\exp\left(-\frac{t^2}{2\sigma^2}\right) \tag{21}$$

for some $\sigma > 0$. Then for any $\delta > 0$, with probability at least $1 - \delta$, we have

$$\left\|\sum_{i=1}^n Z_i\right\|_2 \leq 3\sqrt{n\sigma^2 \log(2d/\delta)}.$$

Proof Given Corollary 7 in Jin et al. (2019), the only thing to prove is that their constant can be set equal to 3. Their Corollary 7 follows from their Lemma 6 in which it is shown that when the matrix Y defined in (20) satisfies

$$\forall \theta \in \mathbb{R}, \qquad \mathbb{E}[\exp(\theta Y)] \preceq \exp(c\theta^2\sigma^2) I,$$

then we have for any $\theta > 0$, with probability at least $(1 - \delta)$,

$$\left\|\sum_{i=1}^n Z_i\right\|_2 \leq c \cdot \theta n\sigma^2 + \frac{1}{\theta}\log(2d/\delta).$$

Taking $\theta = \sqrt{\log(2d/\delta)/(cn\sigma^2)}$ yields

$$\left\|\sum_{i=1}^n Z_i\right\|_2 \leq 2\sqrt{c}\sqrt{n\sigma^2 \log(2d/\delta)},$$

and we conclude with Lemma 5 ($c = 9/4$, $2\sqrt{c} = 3$). □

B Proof of Theorem 1

The proof is organized as follows. We first provide an upper bound on the error (Step 1). This bound involves the norm of the error made on the rescaled coefficients and is controlled in Step 2. Then (Step 3), we construct an event that has probability at least $1 - \delta$ on which we can control the terms that appear in the upper bound of Step 2. Collecting all the inequalities, we will arrive at the stated bound (Step 4).

Step 1. — Since $f = P(f) + \beta^\star(f)^T h + \epsilon$, the oracle estimate of $P(f)$, which uses the unknown, optimal coefficient vector $\beta^\star(f)$, is

$$\hat{\alpha}_n^{\mathrm{or}}(f) = P_n[f - \beta^\star(f)^T h] = P(f) + P_n(\epsilon).$$

The difference between the OLS and oracle estimates is

$$\hat{\alpha}_n^{\mathrm{ols}}(f) - \hat{\alpha}_n^{\mathrm{or}}(f) = \left(\beta^\star(f) - \hat{\beta}_n^{\mathrm{ols}}(f)\right)^T P_n(h).$$

Let $G = P(hh^\top)$ be the $m \times m$ Gram matrix. By assumption, G is positive definite. Write

$$\eta^\star = G^{1/2}\beta^\star(f), \quad \hat{\eta} = G^{1/2}\hat{\beta}_n^{\mathrm{ols}}(f), \quad \hbar = G^{-1/2}h(f).$$

The estimation error of the OLS estimator can thus be decomposed as

$$\begin{aligned} &n\left(\hat{\alpha}_n^{\mathrm{ols}}(f) - P(f)\right) \\ &\quad = n\left(\hat{\alpha}_n^{\mathrm{or}}(f) - P(f)\right) + \left(\beta^\star(f) - \hat{\beta}_n^{\mathrm{ols}}(f)\right)^T n\, P_n(h) \\ &\quad = \sum_{i=1}^n \epsilon(X_i) + \left(\beta^*(f) - \hat{\beta}_n^{\mathrm{ols}}(f)\right)^T \sum_{i=1}^n h(X_i) \\ &\quad = \sum_{i=1}^n \epsilon(X_i) + (\eta^\star - \hat{\eta})^\top \sum_{i=1}^n \hbar(X_i). \end{aligned}$$

By the triangle and Cauchy–Schwarz inequalities,

$$\begin{aligned} n\left|\hat{\alpha}_n^{\mathrm{ols}}(f) - P(f)\right| &\le \left|\textstyle\sum_{i=1}^n \epsilon(X_i)\right| \\ &\quad + \left\|\eta^\star - \hat{\eta}\right\|_2 \left\|\textstyle\sum_{i=1}^n \hbar(X_i)\right\|_2. \end{aligned} \tag{22}$$

Step 2. — We will show that, if $\lambda_{\min}(P_n(\hbar\hbar^\top)) > \|P_n(\hbar)\|_2^2$, then

$$\left\|\hat{\eta} - \eta^\star\right\|_2 \le \frac{\|P_n(\hbar\epsilon)\|_2 + \|P_n(\hbar)\|_2\,|P_n(\epsilon)|}{\lambda_{\min}(P_n(\hbar\hbar^\top)) - \|P_n(\hbar)\|_2^2}. \tag{23}$$

and thus, by (22),

$$\begin{aligned} &\left|\hat{\alpha}_n^{\mathrm{ols}}(f) - P(f)\right| \\ &\quad \le |P_n(\epsilon)| + \frac{\|P_n(\hbar\epsilon)\|_2 + \|P_n(\hbar)\|_2\,|P_n(\epsilon)|}{\lambda_{\min}(P_n(\hbar\hbar^\top)) - \|P_n(\hbar)\|_2^2}\,\|P_n(\hbar)\|_2 \end{aligned} \tag{24}$$

Step 2.1 — Considered the column-centered $n \times m$ design matrices

$$\begin{aligned} H_c &= H - \mathbb{1}_n P_n(h)^\top = \left(h_j(X_i) - P_n(h_j)\right)_{i,j}, \\ \bar{H}_c &= H_c G^{-1/2} = \bar{H} - \mathbb{1}_n P_n(\hbar)^\top = \left(\hbar_j(X_i) - P_n(\hbar_j)\right)_{i,j}. \end{aligned}$$

Since $\bar{H}^\top \mathbb{1}_n = n P_n(\hbar)$, we have

$$\begin{aligned} \bar{H}_c^\top \bar{H}_c &= \bar{H}^\top \bar{H} - n P_n(\hbar) P_n(\hbar)^\top \\ &= n\left(P_n(\hbar\hbar^\top) - P_n(\hbar) P_n(\hbar)^\top\right). \end{aligned}$$

As a consequence, for $u \in \mathbb{R}^m$,

$$\begin{aligned} u^\top \bar{H}_c^\top \bar{H}_c u &= n\left(u^\top P_n(\hbar\hbar^\top)u - (P_n(\hbar)^\top u)^2\right) \\ &\ge n\left(\lambda_{\min}(P_n(\hbar\hbar^\top)) - \|P_n(\hbar)\|_2^2\right)\|u\|_2^2. \end{aligned}$$

by the Cauchy–Schwarz inequality. In particular, $u^\top \bar{H}_c^\top \bar{H}_c u$ is non-zero for non-zero $u \in \mathbb{R}^m$, so that $\bar{H}_c^\top \bar{H}_c$ is invertible, and so is the matrix

$$H_c^\top H_c = G^{1/2} \bar{H}_c \bar{H}_c G^{1/2}.$$

Also, the smallest eigenvalue of $\bar{H}_c^\top \bar{H}_c$ is bounded from below by

$$\lambda_{\min}(\bar{H}_c^\top \bar{H}_c) \ge n\left(\lambda_{\min}(P_n(\hbar\hbar^\top)) - \|P_n(\hbar)\|_2^2\right) > 0.$$

The largest eigenvalue of the inverse matrix $(\bar{H}_c^\top \bar{H}_c)^{-1}$ is then bounded from above by

$$\lambda_{\max}\left((\bar{H}_c^\top \bar{H}_c)^{-1}\right) \le \frac{1}{n\left(\lambda_{\min}(P_n(\hbar\hbar^\top)) - \|P_n(\hbar)\|_2^2\right)}. \tag{25}$$

Step 2.2. — Write $\epsilon_c^{(n)} = (\epsilon(X_i) - P_n(\epsilon))_{i=1}^n$ for the centered vector of error terms. Recall $f_c^{(n)} = (f(X_i) - P_n(f))_{i=1}^n$, the centered vector of samples from the integrand. As $f = P(f) + h^\top \beta^*(f) + \epsilon$, we have

$$f_c^{(n)} = H_c \beta^\star(f) + \epsilon_c^{(n)}.$$

From the characterization (4) of the OLS estimate of the coefficient vector and since $H_c^\top H_c$ is invertible,

$$\begin{aligned} \hat{\beta}_n^{\mathrm{ols}}(f) &= (H_c^\top H_c)^{-1} H_c^\top f_c^{(n)} \\ &= (H_c^\top H_c)^{-1} H_c^\top \left(H_c \beta^\star(f) + \epsilon_c^{(n)}\right) \\ &= \beta^\star(f) + (H_c^\top H_c)^{-1} H_c^\top \epsilon_c^{(n)}. \end{aligned}$$

We obtain

$$\begin{aligned} \hat{\eta} - \eta^\star &= G^{1/2}\left(\hat{\beta}_n^{\mathrm{ols}}(f) - \beta^\star(f)\right) \\ &= G^{1/2}(H_c^\top H_c)^{-1} H_c^\top \epsilon_c^{(n)} \\ &= (\bar{H}_c^\top \bar{H}_c)^{-1} \bar{H}_c^\top \epsilon_c^{(n)}. \end{aligned} \tag{26}$$

Step 2.3. — We combine the results from Steps 2.1 and 2.2. From the upper bound (25) and the identity (26), we obtain

$$\left\|\hat{\eta}-\eta^{\star}\right\|_2 \leq \frac{\left\|\bar{H}_c^{\top}\epsilon_c^{(n)}\right\|_2}{n\left(\lambda_{\min}(P_n(\hbar\hbar^{\top}))-\|P_n(\hbar)\|_2^2\right)}$$

Finally, as $\bar{H}_c=(\hbar_j(X_i)-P_n(\hbar_j))_{i,j}$, we find

$$\begin{aligned} n^{-1}\left\|\bar{H}_c^{\top}\epsilon_c^{(n)}\right\|_2 &= n^{-1}\left\|\sum_{i=1}^n \hbar(X_i)\epsilon(X_i)-P_n(\hbar)\sum_{i=1}^n\epsilon(X_i)\right\|_2 \\ &= \|P_n(\hbar\epsilon)-P_n(\hbar)P_n(\epsilon)\|_2 \\ &\leq \|P_n(\hbar\epsilon)\|_2+\|P_n(\hbar)\|_2\,|P_n(\epsilon)|\,. \end{aligned}$$

Equation (23) follows.

Step 3. — In view of (24), we need to ensure that $|P_n(\epsilon)|$, $\|P_n(\hbar)\|$ and $\|P_n(\hbar\epsilon)\|_2$ are small and that $\lambda_{\min}(P_n(\hbar\hbar^{\top}))$ is large. Let $\delta>0$. We construct an event with probability at least $1-\delta$ on which four inequalities hold simultaneously. Recall $B=\sup_{x\in\mathscr{X}}\|\hbar(x)\|_2^2$, defined in (7).

Step 3.1. — Because $\epsilon\in\mathscr{G}(\tau^2)$, Chernoff's inequality (or Lemma 1 with $p=1$) implies that with probability at least $1-\delta/4$,

$$\left|\sum_{i=1}^n\epsilon(X_i)\right| \leq \sqrt{2n\tau^2\log(8/\delta)}. \tag{27}$$

Step 3.2. — For the term $\left\|\sum_{i=1}^n\hbar(X_i)\right\|_2$, we apply the vector Bernstein bound in (Hsu et al. 2014, Lemma 9). On the one hand $\sup_{x\in\mathscr{X}}\|\hbar(x)\|_2\leq\sqrt{B}$ and on the other hand

$$\sum_{i=1}^n\mathbb{E}[\|\hbar(X_i)\|_2^2]=\sum_{i=1}^n\sum_{j=1}^m P(\hbar_j^2)=nm.$$

The cited vector Bernstein bound gives

$$\forall t\geq 0,\ \mathbb{P}\left[\left\|\sum_{i=1}^n\hbar(X_i)\right\|_2>\sqrt{nm}\left(1+\sqrt{8t}\right)+\tfrac{4}{3}t\sqrt{B}\right]\leq e^{-t}.$$

Setting $t=\log(4/\delta)$, we find that, with probability at least $1-\delta/4$, we have

$$\left\|\sum_{i=1}^n\hbar(X_i)\right\|_2\leq\sqrt{nm}\left(1+\sqrt{8\log(4/\delta)}\right)+\tfrac{4}{3}\log(4/\delta)\sqrt{B}.$$

Since $\log(4/\delta)\geq\log(4)$, we have

$$1+\sqrt{8\log(4/\delta)}\leq 4\sqrt{\log(4/\delta)}$$

and thus

$$\begin{aligned}\left\|\sum_{i=1}^n\hbar(X_i)\right\|_2 &\leq 4\sqrt{nm\log(4/\delta)}+\tfrac{4}{3}\log(4/\delta)\sqrt{B} \\ &= 4\sqrt{\log(4/\delta)}\left(\sqrt{nm}+\tfrac{1}{3}\sqrt{B\log(4/\delta)}\right).\end{aligned}$$

The condition on n easily implies that

$$\tfrac{1}{3}\sqrt{B\log(4/\delta)}\leq\tfrac{1}{4}\sqrt{nm}$$

and thus

$$\left\|\sum_{i=1}^n\hbar(X_i)\right\|_2\leq 5\sqrt{nm\log(4/\delta)}. \tag{28}$$

Step 3.3. — To control $\left\|\sum_{i=1}^n\hbar(X_i)\epsilon(X_i)\right\|_2$, we apply Lemma 6 with $Z_i=\hbar(X_i)\epsilon(X_i)$. The random vectors $\hbar(X_i)\epsilon(X_i)$ for $i=1,\ldots,n$ are independent and identically distributed and have mean zero. Since $\|\hbar(X_i)\|_2\leq\sqrt{B}$ by (7) and since $\epsilon\in\mathscr{G}(\tau^2)$ by Assumption 1, we have, for all $t>0$,

$$\begin{aligned}\mathbb{P}[\|\hbar(X_i)\epsilon(X_i)\|_2>t] &\leq \mathbb{P}[\sqrt{B}\,|\epsilon(X_i)|>t] \\ &\leq 2\exp\left(-\tfrac{t^2}{2B\tau^2}\right),\end{aligned}$$

and (21) holds with $\sigma^2=B\tau^2$. Lemma 6 then implies that, with probability at least $1-\delta/4$ and $c=3$ that

$$\left\|\sum_{i=1}^n\hbar(X_i)\epsilon(X_i)\right\|_2\leq c\sqrt{nB\tau^2\log(8m/\delta)}. \tag{29}$$

Step 3.4. — Recall the $n\times m$ matrix $H=(h_j(X_i))_{i,j}$ and put

$$\bar{H}=HG^{-1/2}=(\hbar_j(X_i))_{i,j}.$$

The empirical Gram matrix of the vector $\hbar=(\hbar_1,\ldots,\hbar_m)^{\top}\in L_2(P)^m$ based on the sample $X_1,\ldots,X_n$ is

$$P_n(\hbar\hbar^{\top})=n^{-1}\bar{H}^T\bar{H}.$$

We apply Lemma 2 with $g=\tilde{g}=\hbar$, $p=m$, and δ replaced by $\delta/4$. We find that, with probability at least $1-\delta/4$,

$$\begin{aligned}\forall u\in\mathbb{R}^m,\ \left\|\bar{H}u\right\|_2^2 &= n\,u^{\top}P_n(\hbar\hbar^{\top})u \\ &\geq n\left(1-\sqrt{2Bn^{-1}\log(4m/\delta)}\right)\|u\|_2^2\,.\end{aligned} \tag{30}$$

Since $P_n(\hbar\hbar^{\top})=n^{-1}\bar{H}^{\top}\bar{H}$, it follows that

$$\lambda_{\min}(P_n(\hbar\hbar^{\top}))\geq 1-\sqrt{2Bn^{-1}\log(4m/\delta)}\geq\tfrac{2}{3} \tag{31}$$

as the assumption on n implies that $2Bn^{-1}\log(4m/\delta)\leq 1/9$.

By the union bound, the inequalities (27), (28), (29), and (30) hold simultaneously on an event with probability at least $1-\delta$. For the remainder of the proof, we work on this event, denoted by E.

Step 4. — We combine the bound (24) on the estimation error with the bounds valid on the event E constructed in Step 3. By (31), we have

$$\lambda_{\min}(P_n(\hbar\hbar^\top)) - \|P_n(\hbar)\|_2^2 \geq \tfrac{2}{3} - 25mn^{-1}\log(4/\delta) \geq \tfrac{1}{3}$$

since the assumption on n implies that $25mn^{-1}\log(4/\delta) \leq 1/3$. As $B \geq m \geq 1$, we have

$$\begin{aligned}
&\|P_n(\hbar\epsilon)\|_2 + \|P_n(\hbar)\|_2\,|P_n(\epsilon)| \\
&\quad\leq c\sqrt{n^{-1}B\tau^2\log(8m/\delta)} \\
&\qquad + 5\sqrt{n^{-1}m\log(4/\delta)}\cdot\sqrt{2n^{-1}\tau^2\log(8/\delta)} \\
&\quad\leq \sqrt{n^{-1}B\tau^2\log(8m/\delta)}\left(c + 5\sqrt{2n^{-1}\log(4/\delta)}\right) \\
&\quad\leq (c+\sqrt{2/3})\sqrt{n^{-1}B\tau^2\log(8m/\delta)},
\end{aligned}$$

since, by assumption, $n \geq 75m\log(4/\delta)$ which implies that $\sqrt{n^{-1}\log(4/\delta)} \leq 1/(5\sqrt{3})$. We find

$$\begin{aligned}
|\hat{\alpha}_n^{\mathrm{ols}}(f) - P(f)| &\leq \sqrt{2\tau^2 n^{-1}\log(8/\delta)} \\
&\quad + \frac{1}{1/3}\cdot(c+\sqrt{2/3})\sqrt{n^{-1}B\tau^2\log(8m/\delta)}\cdot 5\sqrt{mn^{-1}\log(4/\delta)} \\
&= \sqrt{2\tau^2 n^{-1}\log(8/\delta)} \\
&\quad + 15(c+\sqrt{2/3})n^{-1}\sqrt{B\tau^2 m\log(8m/\delta)\log(4/\delta)},
\end{aligned}$$

and the value $c = 3$ gives $15(c+\sqrt{2/3}) \approx 57.2 < 58$ which is the bound stated in Theorem 1. □

C Proof of Theorem 2

For a vector $\beta \in \mathbb{R}^m$ and for a non-empty set $S \subset \{1,\ldots,m\}$, write $\beta_S = (\beta_k)_{k\in S}$. For any matrix $A \in \mathbb{R}^{n\times m}$ and $k \in \{1,\ldots,m\}$, let A_k denote its k-th column and if $S = \{k_1,\ldots,k_\ell\} \subset \{1,\ldots,m\}$ with $k_1 < \ldots < k_\ell$, write $A_S = (A_{k_1},\ldots,A_{k_\ell}) \in \mathbb{R}^{n\times\ell}$.

The proof is organized in a similar way as the one of Theorem 1. We first provide an initial upper bound on the error (Step 1). Then we construct an event that (Step 2) has probability at least $1-\delta$ and (Steps 3, 4, 5) on which we can control each of the terms of the previous upper bound. The combination of all steps to deduce the final statement is made clear in Step 6.

Step 1. — As in the proof of Theorem 1, with $\hat{\beta}_n^{\mathrm{ols}}(f)$ replaced by $\hat{\beta}_n^{\mathrm{lasso}}(f)$, the estimation error of the LASSO estimator can be decomposed as

$$\begin{aligned}
&n\left(\hat{\alpha}_n^{\mathrm{lasso}}(f) - P(f)\right) \\
&\quad= \sum_{i=1}^n \epsilon(X_i) + \left(\beta^*(f) - \hat{\beta}_n^{\mathrm{lasso}}(f)\right)^T \sum_{i=1}^n h(X_i).
\end{aligned}$$

Writing $\hat{u} = \hat{\beta}_n^{\mathrm{lasso}}(f) - \beta^\star(f)$, we get, by the triangle and Hölder inequalities,

$$n\left|\hat{\alpha}_n^{\mathrm{lasso}}(f) - P(f)\right| \leq \left|\sum_{i=1}^n \epsilon(X_i)\right| + \|\hat{u}\|_1 \max_{k=1,\ldots,m}\left|\sum_{i=1}^n h_k(X_i)\right|. \tag{32}$$

Step 2. — Let $\delta > 0$. We construct an event, E, with probability at least $1-\delta$ on which four inequalities, namely (33), (34), (35) and (36), hold simultaneously.

– Since $\epsilon \in \mathscr{G}(\tau^2)$, we can apply Lemma 1 with $p = 1$ to get that, with probability at least $1-\delta/4$,

$$\left|\sum_{i=1}^n \epsilon(X_i)\right| \leq \sqrt{2n\tau^2\log(8/\delta)}. \tag{33}$$

– In view of (Boucheron et al. 2013, Lemma 2.2) and Assumption 2, we have $h_k \in \mathscr{G}(U_h^2)$ for all $k = 1,\ldots,m$. Hence we can apply Lemma 1 with $p = m$ to get that, with probability at least $1-\delta/4$,

$$\max_{k=1,\ldots,m}\left|\sum_{i=1}^n h_k(X_i)\right| \leq \sqrt{2nU_h^2\log(8m/\delta)}. \tag{34}$$

– By virtue of Assumptions 1 and 2, we can apply Lemma 4 to find $h_k\epsilon \in \mathscr{G}(C\tau^2U_h^2)$ with $C = 9/2$. Hence we can apply Lemma 1 to get that, with probability at least $1-\delta/4$,

$$\max_{k=1,\ldots,m}\left|\sum_{i=1}^n h_k(X_i)\epsilon(X_i)\right| \leq \sqrt{2nC\tau^2U_h^2\log(8m/\delta))}. \tag{35}$$

– In view of (Boucheron et al. 2013, Lemma 2.2) and Assumptions 2 and 6, we have $h_kh_l - P(h_kh_l) \in \mathscr{G}(U_h^4)$ for all $k,l \in \{1,\ldots,m\}$. Hence we can apply Lemma 1 with $p = m^2$ to get that, with probability at least $1-\delta/4$,

$$\max_{\substack{1\leq k\leq m\\ 1\leq l\leq m}}\left|\sum_{i=1}^n\{h_k(X_i)h_l(X_i) - P(h_kh_l)\}\right| \leq \sqrt{2nU_h^4\log(8m^2/\delta)}.$$

Denote by $\Delta = (P_n - P)\{hh^T\}$. Because by assumption $2(\ell^\star/\gamma^\star)\sqrt{2U_h^4\log(8m^2/\delta)} \leq \sqrt{n}$, we have that

$$(\ell^\star/\gamma^\star)\max_{1\leq k,l\leq m}|\Delta_{k,l}| \leq 1/2.$$

Remark that

$$\forall u \in \mathbb{R}^m, \quad n^{-1} \|Hu\|_2^2 - u^T G u = u^T \Delta u.$$

Then, following (Bickel et al. 2009, equation (3.3)), use the inequality $|u^T \Delta u| \le \|u\|_1^2 \max_{1\le k,l\le m} |\Delta_{k,l}|$, to obtain that, with probability $1-\delta/4$, for all $u \in \mathscr{C}(S^\star; 3)$,

$$\begin{aligned}\|Hu\|_2^2/n &\ge u^T G u - \|u\|_1^2 \max_{1\le k,l\le m} |\Delta_{k,l}| \\ &\ge u^T G u - \|u\|_2^2 \ell^\star \max_{1\le k,l\le m} |\Delta_{k,l}| \\ &\ge u^T G u - (u^T G u)(\ell^\star/\gamma^\star) \max_{1\le k,l\le m} |\Delta_{k,l}| \\ &\ge (u^T G u)/2.\end{aligned}$$

It follows that with probability at least $1-\delta/4$,

$$\|Hu\|_2^2 \ge (n\gamma^\star/2)\, \|u\|_2^2 . \tag{36}$$

Step 3. — We claim that, on the event E, we have

$$\forall u \in \mathscr{C}(S^\star; 3), \quad \|H_c u\|_2^2 \ge (n\gamma^\star/4)\, \|u\|_2^2 \tag{37}$$

We have

$$H_c^T H_c = H^T H - n\, P_n(h)\, P_n(h)^T$$

and thus,

$$\|H_c u\|_2^2 \ge \|Hu\|_2^2 - n \max_{k=1,\dots,m} |P_n(h_k)|^2\, \|u\|_1^2 .$$

We treat both terms on the right-hand side. On the one hand, we just have obtained a lower bound for the first term. On the other hand, in view of (34) and because $\|u\|_1^2 \le 16\, \|u_{S^\star}\|_1^2 \le 16\ell^\star\, \|u\|_2^2$, we have

$$\begin{aligned}\|u\|_1^2 \max_{k=1,\dots,m} |P_n(h_k)|^2 &= \|u\|_1^2\, n^{-2} \cdot \max_{k\in S^\star} \Big|\sum_{i=1}^n h_k(X_i)\Big|^2 \\ &\le 16\ell^\star\, \|u\|_2^2 \cdot n^{-2} \cdot 2nU_h^2 \log(8m/\delta) \\ &\le \|u\|_2^2\, \gamma^\star/4\end{aligned}$$

as $n \ge (16\times 8)\ell^\star(U_h^2/\gamma^\star)\log(8m/\delta)$ by assumption. In combination with (36), we find

$$\|H_c u\|_2^2 \ge n(\gamma^\star/2)\, \|u\|_2^2 - n(\gamma^\star/4)\, \|u\|_2^2 = n(\gamma^\star/4)\, \|u\|_2^2 .$$

Step 4. — We claim that, on the event E, we have

$$\left\| H_c^T \epsilon_c^{(n)} \right\|_\infty \le (3+\sqrt{2}/8)\sqrt{\log(8m/\delta)}\, U_h \tau \sqrt{n}. \tag{38}$$

Indeed, on the left-hand side in (38) we have in virtue of (33), (34) and (35),

$$\begin{aligned}&\left\| H_c^T \epsilon_c^{(n)} \right\|_\infty \\ &= \max_{k=1,\dots,m} \left| \sum_{i=1}^n (h_k(X_i) - P_n(h_k))(\epsilon(X_i) - P_n(\epsilon)) \right| \\ &= \max_{k=1,\dots,m} \left| \left(\sum_{i=1}^n h_k(X_i)\epsilon(X_i) \right) - n P_n(h_k) P_n(\epsilon) \right| \\ &\le \max_{k=1,\dots,m} \left| \sum_{i=1}^n h_k(X_i)\epsilon(X_i) \right| \\ &\quad + n^{-1} \left| \sum_{i=1}^n \epsilon(X_i) \right| \max_{k=1,\dots,m} \left| \sum_{i=1}^n h_k(X_i) \right| \\ &\le \sqrt{2nC\tau^2 U_h^2 \log(8m/\delta)} \\ &\quad + n^{-1} \sqrt{2n\tau^2 \log(8/\delta)} \sqrt{2nU_h^2 \log(8m/\delta)} \\ &= \sqrt{2nC\tau^2 U_h^2 \log(8m/\delta)} \left(1 + \sqrt{2\log(8/\delta)/(Cn)} \right).\end{aligned}$$

Since $\ell^\star \ge 1$ and $\ell^\star U_h^2 \ge \sum_{k\in S^\star} P(h_k^2) \ge \gamma^\star$, the assumed lower bound on n implies that $n \ge 128 \log(8/\delta)$. As $C = 9/2$, the factor $\sqrt{2C}(1+\sqrt{2\log(8/\delta)/(Cn)})$ is bounded by $3+\sqrt{2}/8$ and we get (38).

Step 5. — Recall $\hat{u} = \hat{\beta}_n^{\text{lasso}}(f) - \beta^\star(f)$. We claim that, on the event E, we have

$$\left\| \hat{u} \right\|_1 \le 48\lambda\ell^\star/\gamma^\star. \tag{39}$$

To prove this result, we shall rely on the following lemma.

Lemma 7 *If* $n\lambda \ge 2 \left\| H_c^T \epsilon_c^{(n)} \right\|_\infty$ *then, writing* $\hat{u} = \hat{\beta}_n^{\text{lasso}}(f) - \beta^\star(f)$, *we have* $\hat{u} \in \mathscr{C}(S^\star; 3)$ *and*

$$\left\| H_c \hat{u} \right\|_2^2 \le 3n\lambda \left\| \hat{u}_{S^\star} \right\|_1 . \tag{40}$$

Proof This is just a reformulation of the reasoning on p. 298 in (Tibshirani et al. 2015) with a slightly sharper upper bound. The vector $\hat{v}$ at the right-hand side of their Eq. (11.23) can be replaced by $\hat{v}_S$. For the sake of completeness, we provide the details.

In the proof we use the short-cuts $\beta^\star = \beta^\star(f)$ and $\hat{\beta}_n^{\text{lasso}} = \hat{\beta}_n^{\text{lasso}}(f)$. Recall $\epsilon_c^{(n)} = f_c^{(n)} - H_c \beta^\star(f)$ and define

$$\begin{aligned}G(u) &= \|f_c^{(n)} - H_c(\beta^\star + u)\|_2^2/(2n) + \lambda\|\beta^\star + u\|_1 \\ &= \|\epsilon_c^{(n)} - H_c u\|_2^2/(2n) + \lambda\|\beta^\star + u\|_1 .\end{aligned}$$

Because $G(\hat{u}) \le G(0)$, we have

$$\|H_c\hat{u}\|_2^2/(2n) \le \hat{u}^T H_c^T \epsilon_c^{(n)}/n + \lambda(\|\beta^\star\|_1 - \|\beta^\star + \hat{u}\|_1)$$

From the triangle inequality

$$\|(\beta^\star - (-\hat{u}))_{S^\star}\|_1 \geq |\|\beta^\star_{S^\star}\|_1 - \|\hat{u}_{S^\star}\|_1| \geq \|\beta^\star_{S^\star}\|_1 - \|\hat{u}_{S^\star}\|_1,$$

implying that

$$\begin{aligned}
&\|\beta^\star\|_1 - \|\beta^\star + \hat{u}\|_1 \\
&\quad = \|\beta^\star\|_1 - \|(\beta^\star + \hat{u})_{S^\star}\|_1 - \|(\beta^\star + \hat{u})_{\overline{S^\star}}\|_1 \\
&\quad \leq \|\beta^\star\|_1 - \|\beta^\star_{S^\star}\|_1 + \|\hat{u}_{S^\star}\|_1 - \|(\beta^\star + \hat{u})_{\overline{S^\star}}\|_1 \\
&\quad = \|\hat{u}_{S^\star}\|_1 - \|\hat{u}_{\overline{S^\star}}\|_1.
\end{aligned}$$

From Hölder's inequality, we get

$$\left|\hat{u}^T H_c^T \epsilon_c^{(n)}\right| \leq \left\|H_c^T \epsilon_c^{(n)}\right\|_\infty \cdot \left\|\hat{u}\right\|_1,$$

which leads to

$$\left\|H_c \hat{u}\right\|_2^2/(2n) \leq \|H_c^T \epsilon_c^{(n)}\|_\infty \|\hat{u}\|_1/n + \lambda(\left\|\hat{u}_{S^\star}\right\|_1 - \left\|\hat{u}_{\overline{S^\star}}\right\|_1).$$

Consequently, because $\left\|H_c^T \epsilon_c^{(n)}\right\|_\infty /n \leq \lambda/2$ by assumption, we obtain

$$\begin{aligned}
0 \leq \|H_c \hat{u}\|_2^2/(2n) &\leq \lambda(\|\hat{u}\|_1/2 + \|\hat{u}_{S^\star}\|_1 - \|\hat{u}_{\overline{S^\star}}\|_1) \\
&= (\lambda/2)(3\|\hat{u}_{S^\star}\|_1 - \|\hat{u}_{\overline{S^\star}}\|_1).
\end{aligned}$$

The right-hand side must be nonnegative, whence $\left\|\hat{u}_{\overline{S^\star}}\right\|_1 \leq 3\left\|\hat{u}_{S^\star}\right\|_1$, i.e., $\hat{u} \in \mathscr{C}(S; 3)$. The bound in (40) follows as well. □

On the event E, the conclusion of Lemma 7 is valid because the bound on $\left\|H_c^T \epsilon_c^{(n)}\right\|_\infty$ in (38) and the assumption on λ in Theorem 2 together imply that $\lambda \geq 2\left\|H_c^T \epsilon_c^{(n)}\right\|_\infty /n$. The cone property of Lemma 7 yields $\hat{u} \in \mathscr{C}(S^\star; 3)$ so that

$$\left\|\hat{u}\right\|_1 = \left\|\hat{u}_{S^\star}\right\|_1 + \left\|\hat{u}_{\overline{S^\star}}\right\|_1 \leq 4\left\|\hat{u}_{S^\star}\right\|_1. \tag{41}$$

Thanks to (37) and Lemma 7, and since $|S^\star| = \ell^\star$, we get

$$\begin{aligned}
\left\|\hat{u}_{S^\star}\right\|_1^2 &\leq \ell^\star \left\|\hat{u}_{S^\star}\right\|_2^2 \\
&\leq \ell^\star \left\|\hat{u}\right\|_2^2 \\
&\leq \ell^\star \cdot n^{-1}(4/\gamma^\star)\left\|H_c \hat{u}\right\|_2^2 \\
&\leq \ell^\star \cdot n^{-1}(4/\gamma^\star) \cdot 3n\lambda \left\|\hat{u}_{S^\star}\right\|_1 = 12\ell^\star(\lambda/\gamma^\star)\left\|\hat{u}_{S^\star}\right\|_1.
\end{aligned}$$

It follows that $\left\|\hat{u}_{S^\star}\right\|_1 \leq 12\ell^\star\lambda/\gamma^\star$. In combination with (41), we find (39).

Step 6. — Equation (32) gave a bound on the estimation error involving three terms. On the event E, these terms were shown to be bounded in (33), (34), and (39). It follows that, on E, we finally have

$$\begin{aligned}
&n\left|\hat{\alpha}_n^{\text{lasso}}(f) - P(f)\right| \\
&\quad \leq \sqrt{2n\tau^2 \log(8/\delta)} + 48\lambda\ell^\star/\gamma^\star \cdot \sqrt{2nU_h^2 \log(8m/\delta)}.
\end{aligned}$$

Divide by n and use $48\sqrt{2} < 68$ to obtain (9). □

D Proof of Theorem 3

Recall that $S^\star = \{j = 1, \ldots, m : \beta_j^\star(f) \neq 0\}$ with $\ell^\star = |S^\star|$ and that $\overline{S^\star} = \{1, \ldots, m\} \setminus S^\star$. Further, $H_{c,S^\star}$ is the $n \times \ell^\star$ matrix having columns $H_{c,k}$ for $k \in S^\star$, where $H_{c,k}$ is the k-th column of H_c.

Step 1. — We first establish some (non-probabilistic) properties of $\hat{\beta}_n^{\text{lasso}}(f)$. To this end, we consider the linear regression of the non-active control variates on the active ones: for $k \in \overline{S^\star} = \{j = 1, \ldots, m : \beta_j^\star(f) = 0\}$, this produces the coefficient vector

$$\hat{\theta}_n^{(k)} \in \arg\min_{\theta \in \mathbb{R}^{\ell^\star}} \left\|H_{c,k} - H_{c,S^\star}\theta\right\|_2.$$

Further, we consider the OLS oracle estimate $\hat{\beta}_n^\star$, which is the OLS estimator based upon the active control variables only, i.e.,

$$\hat{\beta}_n^\star \in \arg\min_{\beta \in \mathbb{R}^{\ell^\star}} \|f_c^{(n)} - H_{c,S^\star}\beta\|_2.$$

Our assumptions will imply that, with large probability, $H_{c,S^\star}$ has rank ℓ^*, in which case

$$\begin{aligned}
\hat{\theta}_n^{(k)} &= (H_{c,S^\star}^T H_{c,S^\star})^{-1} H_{c,S^\star}^T H_{c,k}, \\
\hat{\beta}_n^\star &= (H_{c,S^\star}^T H_{c,S^\star})^{-1} H_{c,S^\star}^T f_c^{(n)}.
\end{aligned}$$

The following lemma provides a number of (non-probabilistic) properties of $\hat{\beta}_n^{\text{lasso}}(f)$, given certain conditions on H_c and $\epsilon_c^{(n)}$. Recall that a norm $\|\cdot\|$ on $\mathbb{R}^p$ induces a matrix norm on $\mathbb{R}^{p\times p}$ via $\|A\| = \sup\{\|Au\| : u \in \mathbb{R}^p, \|u\| = 1\}$ for $A \in \mathbb{R}^{p\times p}$.

Lemma 8 *If $H_{c,S^\star}$ has rank $\ell^\star$ and if there exists $\kappa \in (0, 1]$ such that*

$$\max_{k \in \overline{S^\star}} \left\|\hat{\theta}_n^{(k)}\right\|_1 \leq 1 - \kappa, \tag{42}$$

$$\max_{k \in \overline{S^\star}} \left|(H_{c,k} - H_{c,S^\star}\hat{\theta}_n^{(k)})^T \epsilon_c^{(n)}\right| \leq \kappa\lambda n, \tag{43}$$

then the minimizer $\hat{\beta}_n^{\text{lasso}}(f)$ in (5) is unique, with support $\text{supp}(\hat{\beta}_n^{\text{lasso}}(f)) \subset S^\star$, and it satisfies

$$\max_{k\in S^\star}\left|\hat\beta^{\text{lasso}}_{n,k}(f)-\beta^\star_k(f)\right| \le \max_{k\in S^\star}\left|\hat\beta^\star_{n,k}-\beta^\star_k(f)\right| + n\lambda\left\|(H^T_{c,S^\star}H_{c,S^\star})^{-1}\right\|_\infty. \quad (44)$$

Proof The proof of the previous result is actually contained in Tibshirani et al. (2015). The uniqueness of the LASSO solution and the property that it does not select inactive covariates follows directly from the proof of their Theorem 11.3. The only difference is that, in our case, the inequality (43) is an assumption whereas in Tibshirani et al. (2015) it is a property of the Gaussian fixed design model. The approach in Tibshirani et al. (2015) is based upon checking the *strict dual feasibility condition*. The bound (44) is Eq. (11.37) in Tibshirani et al. (2015). □

We slightly modify Lemma 8 to make the conditions (42) and (43) easier to check and to make the bound (44) easier to use.

Lemma 9 *If there exists* $\nu>0$ *such that*

$$\forall u\in\mathbb{R}^{\ell^\star},\qquad \left\|H_{c,S^\star}u\right\|_2^2\ge n\nu\left\|u\right\|_2^2, \quad (45)$$

and if there exists $\kappa\in(0,1]$ *such that*

$$\frac{\ell^\star}{\nu n}\max_{k\in\overline{S^\star}}\max_{j\in S^\star}\left|H^T_{c,j}H_{c,k}\right|\le 1-\kappa, \quad (46)$$

$$\max_{k=1,\dots,m}\left|H^T_{c,k}\epsilon^{(n)}_c\right|\le\frac{1}{2}\kappa\lambda n, \quad (47)$$

then the minimizer $\hat\beta^{\text{lasso}}_n(f)$ *in* (5) *is unique, with support satisfying* $\text{supp}(\hat\beta^{\text{lasso}}_n(f))\subset S^\star$, *and it holds that*

$$\max_{k\in S^\star}\left|\hat\beta^{\text{lasso}}_{n,k}(f)-\beta^\star_k(f)\right|\le(1+\kappa/2)\sqrt{\ell^\star}\lambda/\nu. \quad (48)$$

Proof By (45), the smallest eigenvalue of the $\ell^\star\times\ell^\star$ matrix $H^T_{c,S^\star}H_{c,S^\star}$ is positive, so that it is invertible and $H_{c,S^\star}$ has rank $\ell^\star$.

We show that (46) implies (42). For each $k\in\overline{S^\star}$, the vector $\hat\theta^{(k)}_n$ has length $\ell^\star$, so that

$$\left\|\hat\theta^{(k)}_n\right\|_1\le\sqrt{\ell^\star}\left\|\hat\theta^{(k)}_n\right\|_2.$$

Because $\hat\theta^{(k)}_n$ is an OLS estimate, using that the largest eigenvalue of $(H^T_{c,S^\star}H_{c,S^\star})^{-1}$ being bounded from above by $(n\nu)^{-1}$, we obtain

$$\left\|\hat\theta^{(k)}_n\right\|_2=\left\|(H^T_{c,S^\star}H_{c,S^\star})^{-1}H^T_{c,S^\star}H_{c,k}\right\|_2\le\frac{1}{n\nu}\left\|H^T_{c,S^\star}H_{c,k}\right\|_2$$

Since $\|x\|_2\le\sqrt{m}\,\|x\|_\infty$ for $x\in\mathbb{R}^m$, we can conclude that

$$\left\|\hat\theta^{(k)}_n\right\|_2\le\frac{\sqrt{\ell^\star}}{\nu n}\max_{j\in S^\star}\left|H^T_{c,j}H_{c,k}\right|.$$

Combining the two bounds, we find that (46) indeed implies (42).

Next we show that (47) implies (43). For $k\in\overline{S^\star}$, we have

$$\begin{aligned}\left|(H_{c,k}-H_{c,S^\star}\hat\theta^{(k)}_n)^T\epsilon^{(n)}_c\right| &\le\left|H^T_{c,k}\epsilon^{(n)}_c\right|+\left|(\hat\theta^{(k)}_n)^TH^T_{c,S^\star}\epsilon^{(n)}_c\right|\\ &\le\left|H^T_{c,k}\epsilon^{(n)}_c\right|+\left\|\hat\theta^{(k)}_n\right\|_1\max_{j\in S^\star}\left|H^T_{c,j}\epsilon^{(n)}_c\right|.\end{aligned}$$

Using (42) and (47) we deduce (43).

The conditions of Lemma 8 have been verified, and so its conclusion holds. We simplify the two terms in the upper bound (44). First, we use that

$$\left\|\hat\beta^\star_n-\beta^\star(f)\right\|_2=\left\|(H^T_{c,S^\star}H_{c,S^\star})^{-1}H^T_{c,S^\star}\epsilon^{(n)}_c\right\|_2\le\frac{\sqrt{\ell^\star}}{\nu n}\left\|H^T_c\epsilon^{(n)}_c\right\|_\infty.$$

Second, for any matrix $A\in\mathbb{R}^{p\times p}$, we have $\|A\|_\infty\le\sqrt{p}\,\|A\|_2$ (e.g., (Horn and Johnson 2012, page 365)), and this we apply to $(H^T_{c,S^\star}H_{c,S^\star})^{-1}$. In this way, the upper bound in (44) is dominated by

$$\left\|\hat\beta^\star_n-\beta^\star(f)\right\|_2+n\lambda\cdot\sqrt{\ell^\star}\left\|(H^T_{c,S^\star}H_{c,S^\star})^{-1}\right\|_2\le\frac{\sqrt{\ell^\star}}{n\nu}\max_{k\in S^\star}\left|H^T_{c,k}\epsilon^{(n)}_c\right|+n\lambda\cdot\sqrt{\ell^\star}\cdot\frac{1}{n\nu},$$

since the largest eigenvalue of $(H^T_{c,S^\star}H_{c,S^\star})^{-1}$ is at most $(n\nu)^{-1}$. Use (47) to further simplify the right-hand side, yielding (48). □

Step 2. — Let $\delta\in(0,1)$ and $n=1,2,\dots$. In a similar way as in the proof of Theorem 1, we construct an event of probability at least $1-\delta$. This time, we need five inequalities to hold simultaneously.

– Because $\epsilon\in\mathcal{G}(\tau^2)$, with probability at least $1-\delta/5$,

$$\left|\sum_{i=1}^n\epsilon(X_i)\right|\le\sqrt{2n\tau^2\log(10/\delta)}. \quad (49)$$

– In view of (Boucheron et al. 2013, Lemma 2.2) and Assumption 2, we have $h_k\in\mathcal{G}(U_h^2)$ for all $k=1,\dots,m$. Hence we can apply Lemma 1 with $p=m$ to get that, with probability at least $1-\delta/5$,

$$\max_{k=1,\dots,m}\left|\sum_{i=1}^n h_k(X_i)\right|\le\sqrt{2nU_h^2\log(10m/\delta)}. \quad (50)$$

– By virtue of Assumptions 1 and 2, we can apply Lemma 4 to have $h_k\epsilon \in \mathscr{G}(CU_h^2\tau^2)$, where $C = 9/2$. Hence we can apply Lemma 1 to get that, with probability at least $1-\delta/5$,

$$\max_{k=1,\dots,m}\left|\sum_{i=1}^n h_k(X_i)\epsilon(X_i)\right| \le \sqrt{2Cn\tau^2U_h^2\log(10m/\delta))}. \tag{51}$$

– Recall that $B^\star = \sup_{x\in\mathscr{X}} h_{S^\star}^T(x)G_{S^\star}^{-1}h_{S^\star}(x)$ with

$$B^\star \le \lambda_{\max}(G_{S^\star}^{-1}) \sup_{x\in\mathscr{X}} h_{S^\star}^T(x)h_{S^\star}(x) \le \ell^\star U_h^2/\gamma^{\star\star},$$

The assumption on n easily implies that $n \ge 8B^\star\log(5\ell^\star/\delta)$. Applying Lemma 2 with $p = \ell^\star$, $g = h_{S^\star}$, and δ replaced by $\delta/5$, we find that, with probability at least $1-\delta/5$,

$$\|H_{S^\star}u\|_2^2 \ge n\gamma^{\star\star}\|u\|_2^2/2, \qquad \forall u\in\mathbb{R}^{\ell^*}. \tag{52}$$

– Finally, because $\left|h_j(x)\right| \le U_h$ for all $x\in\mathscr{X}$ and $j\in\{1,\dots,m\}$ and because $P(h_kh_j) = 0$ for all $(k,j)\in\overline{S^\star}\times S^\star$, we have $h_kh_j\in\mathscr{G}(U_h^4)$ for such k and j, and thus, with probability at least $1-\delta/5$,

$$\max_{k\in\overline{S^\star}}\max_{j\in S^\star}\left|\sum_{i=1}^n h_k(X_i)h_j(X_i)\right| \le \sqrt{2nU_h^4\log(10\ell^\star m/\delta)}. \tag{53}$$

By the union bound, the event, say E, on which (49), (50), (51), (52) and (53) are satisfied simultaneously has probability at least $1-\delta$. We work on the event E for the rest of the proof.

Step 3. — On the event E, we have

$$\forall u\in\mathbb{R}^{\ell^\star}, \qquad \left\|H_{c,S^\star}u\right\|_2^2 \ge n\alpha\gamma^{\star\star}\|u\|_2^2, \tag{54}$$

where $\alpha\in(0,1/2)$ is an absolute constant whose value will be fixed in Step 6(ii). We have

$$H_{c,S^\star}^TH_{c,S^\star} = H_{S^\star}^TH_{S^\star} - n\,P_n(h_{S^\star})\,P_n(h_{S^\star})^T$$

and thus, by the Cauchy–Schwarz inequality and by (52),

$$\begin{aligned}\left\|H_{c,S^\star}u\right\|_2^2 &\ge \|H_{S^\star}u\|_2^2 - n\,\|P_n(h_{S^\star})\|_2^2\,\|u\|_2^2\\ &\ge n\left(\gamma^{\star\star}/2 - \|P_n(h_{S^\star})\|_2^2\right)\|u\|_2^2.\end{aligned}$$

In view of (50), we have

$$\|P_n(h_{S^\star})\|_2^2 \le \frac{\ell^\star}{n^2}2nU_h^2\log(10m/\delta) = 2\ell^\star\log(10m/\delta)U_h^2/n.$$

We thus get

$$\left\|H_{c,S^\star}u\right\|_2^2 \ge n\gamma^{\star\star}\left[\frac{1}{2} - \frac{2\ell^\star\log(10m/\delta)U_h^2/\gamma^{\star\star}}{n}\right]\|u\|_2^2$$

A sufficient condition for (54) is thus that the term in square brackets is at least α, i.e.,

$$n \ge \frac{2}{1/2-\alpha}\ell^\star\log(10m/\delta)U_h^2/\gamma^{\star\star}$$

Since $\ell^\star\ge 1$ and $U_h^2\ge\gamma^{\star\star}$, a condition of the form

$$n \ge \rho\log(10\ell^\star m/\delta)[\ell^\star(U_h^2/\gamma^{\star\star})]^2 \tag{55}$$

is thus sufficient, with much to spare, provided $\rho > 2/(1/2-\alpha)$. In Step 6(ii), we will choose α in such a way that the constant $\rho = 70$ appearing in the statement of the theorem is sufficient.

Step 4. — On the event E, we have

$$\begin{aligned}&\max_{k\in\overline{S^\star}}\max_{j\in S^\star}|H_{c,j}^TH_{c,k}|\\ &\quad\le \sqrt{2nU_h^4\log(10\ell^\star m/\delta)} + 2U_h^2\log(10m/\delta).\end{aligned} \tag{56}$$

Indeed, denote $A = \overline{S^\star}\times S^\star$, in virtue of (50) and (53), the left-hand side is bounded by

$$\begin{aligned}&\max_{(k,j)\in A}\left|\left(\sum_{i=1}^n h_k(X_i)h_j(X_i)\right) - n\,P_n(h_k)\,P_n(h_j)\right|\\ &\le \max_{(k,j)\in A}\left|\sum_{i=1}^n h_k(X_i)h_j(X_i)\right| + \frac{1}{n}\max_{k\in\overline{S^\star}}\left|\sum_{i=1}^n h_k(X_i)\right|\max_{j\in S^\star}\left|\sum_{i=1}^n h_j(X_i)\right|\\ &\le \max_{(k,j)\in A}\left|\sum_{i=1}^n h_k(X_i)h_j(X_i)\right| + \frac{1}{n}\max_{k=1,\dots m}\left|\sum_{i=1}^n h_k(X_i)\right|^2\\ &\le \sqrt{2nU_h^4\log(10\ell^\star m/\delta)} + \frac{1}{n}2nU_h^2\log(10m/\delta),\end{aligned}$$

which is (56).

Step 5. — On the event E, we have

$$\begin{aligned}&\left\|H_c^T\epsilon_c^{(n)}\right\|_\infty\\ &\quad\le \sqrt{2nC\tau^2U_h^2\log(10m/\delta)}\left(1+\sqrt{2\log(10/\delta)/(Cn)}\right).\end{aligned} \tag{57}$$

The proof is the same as the first part of the one (38).

Step 6. — We will verify that on the event E, the three assumptions of Lemma 9 are satisfied with $\kappa = 1/2$ and $\nu = \alpha\gamma^{\star\star}$, with α as in Step 3.

(i) Eq. (45) with $\nu = \alpha\gamma^{\star\star}$ is just (54).

(ii) Eq. (46) with $\nu = \alpha\gamma^{\star\star}$ and $\kappa = 1/2$ follows from (56) provided we have

$$\frac{\ell^\star}{\alpha\gamma^{\star\star}n}\left(\sqrt{2nU_h^4\log(10\ell^\star m/\delta)} + 2U_h^2\log(10m/\delta)\right) \le 1 - \frac{1}{2}.$$

To check whether this is satisfied, we will make use of the elementary inequality[2]

$$\forall (a,b,c) \in (0,\infty)^3,\ \forall x \ge \sqrt{b^2+4ac}/a, \quad ax^2 \ge bx + c.$$

with $x = \sqrt{n}$ and

$$a = \alpha\gamma^{\star\star}/(2\ell^\star),$$
$$b = \sqrt{2U_h^4\log(10\ell^\star m/\delta)},$$
$$c = 2U_h^2\log(10m/\delta).$$

Sufficient is that $n = x^2$ is bounded from below by $(b^2+4ac)/a^2 = (b/a)^2 + 4c/a$, which is

$$\frac{2U_h^4\log(10\ell^\star m/\delta)}{(\alpha\gamma^{\star\star}/(2\ell^\star))^2} + 4\frac{2U_h^2\log(10m/\delta)}{\alpha\gamma^{\star\star}/(2\ell^\star)} = \frac{8}{\alpha^2}\log(10\ell^\star m/\delta)\left(\frac{\ell^\star U_h^2}{\gamma^{\star\star}}\right)^2 + \frac{16}{\alpha}\log(10m/\delta)\left(\frac{\ell^\star U_h^2}{\gamma^{\star\star}}\right).$$

But $\ell^\star \ge 1$ and $\gamma^{\star\star} \le (1/\ell^\star)\sum_{j\in S^\star} P(h_j^2) \le U_h^2$, so that a sufficient condition is that

$$n \ge \left(\frac{8}{\alpha^2} + \frac{16}{\alpha}\right)\log(10\ell^\star m/\delta)[\ell^\star(U_h^2/\gamma^{\star\star})]^2.$$

The constant ρ in (55) must thus be such that

$$\rho \ge \max\left(\frac{2}{1/2-\alpha}, \frac{8}{\alpha^2} + \frac{16}{\alpha}\right).$$

The minimum of the right-hand side as a function of $\alpha \in (0, 1/2)$ occurs at $\alpha = \sqrt{2}/3$ and is equal to $2/(1/2 - \sqrt{2}/3) \approx 69.9$. Taking $\rho = 70$ as in the assumption on n is thus sufficient.

(iii) Eq. (47) with $\kappa = 1/2$ follows from (57), since

$$\sqrt{n\tau^2U_h^2\log(10m/\delta)}\left(\sqrt{2C} + 2\sqrt{\log(10/\delta)/n}\right) \le \lambda n/4$$

by the assumed lower bound on λ. Indeed, since $\ell^\star \ge 1$ and $U_h^2 \ge \gamma^{\star\star}$, the assumed lower bounds on n imply that $n \ge 70\log(10/\delta)$, so that $2\sqrt{\log(10/\delta)/n}$ is bounded by $2/\sqrt{70}$; recall $C = 9/2$. Since $4\cdot(3 + 2/\sqrt{70}) \approx 12.9$, the assumed lower bound for λ suffices.

Step 7. — By the previous step, the conclusions of Lemma 9 with $\kappa = 1/2$ and $\nu = \alpha\gamma^{\star\star}$ hold on the event E, where $\alpha = \sqrt{2}/3$ was specified in Step 6(ii). The minimizer $\hat\beta_n^{\text{lasso}}$ in (5) is thus unique and we have $\text{supp}(\hat\beta_n^{\text{lasso}}(f)) \subset S^\star$.

To show the reverse inclusion, we need to verify that $\left|\hat\beta_{n,k}^{\text{lasso}}(f)\right| > 0$ for all $k \in S^\star$. To this end, we apply (48) with $\kappa = 1/2$ and $\nu = \alpha\gamma^{\star\star}$, which becomes

$$\max_{k\in S^\star}\left|\hat\beta_{n,k}^{\text{lasso}}(f) - \beta_k^\star(f)\right| \le (5/4)\sqrt{\ell^\star}\lambda/(\alpha\gamma^{\star\star}).$$

For any $k \in S^\star$, we thus have

$$\left|\hat\beta_{n,k}^{\text{lasso}}(f)\right| \ge \min_{j\in S^\star}\left|\beta_j^\star(f)\right| - (5/(4\alpha))\sqrt{\ell^\star}\lambda/\gamma^{\star\star}.$$

But for $\alpha = \sqrt{2}/3$, we have approximately $5/(4\alpha) \approx 2.65$. Since $\min_{j\in S^\star}\left|\beta_j^\star(f)\right| > 3\sqrt{\ell^\star}\lambda/\gamma^{\star\star}$ by the assumed upper bound for λ, we find $\left|\hat\beta_{n,k}^{\text{lasso}}(f)\right| > 0$, as required. □

E Proof of Theorem 4

Recall that the LSLASSO estimator is defined as an OLS estimate computed on the active variables selected by the LASSO based on a subsample of size $N \in \{1,\dots,n\}$. Let $\hat\beta_N^{\text{lasso}}(f)$ denote the LASSO coefficient vector in (5) based on the subsample $X_1,\dots,X_N$ and let

$$\hat S_N = \text{supp}(\hat\beta_N^{\text{lasso}}(f)) = \{k \in \{1,\dots,m\} : \hat\beta_{N,k}^{\text{lasso}}(f) > 0\}$$

denote the estimated active set of $\hat\ell = |\hat S_N|$ control variates. The LSLASSO estimate $\hat\alpha_n^{\text{lslasso}}(f)$ based on the full sample $X_1,\dots,X_n$ is defined as the OLS estimator based on the control variates h_k for $k \in \hat S_N$: writing $H_{\hat S_N}$ for the $n \times \hat\ell$ matrix with columns $(h_k(X_i))_{i=1}^n$ with $k \in \hat S_N$, we have

$$\left(\hat\alpha_n^{\text{lslasso}}(f), \hat\beta_n^{\text{lslasso}}(f)\right) \in \operatorname*{arg\,min}_{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^{\hat\ell}}\left\|f^{(n)} - \alpha\mathbb{1}_n - H_{\hat S_N}\beta\right\|_2^2,$$

Therefore, we can derive a concentration inequality by combining the support recovery property (Theorem 3) along with the concentration inequality for the OLS estimate (Theorem 1) using only the active control variates.

[2] The convex parabola $x \mapsto ax^2 - bx - c$ has zeroes at $x_\pm = (b \pm \sqrt{b^2+4ac})/(2a)$, and $x_- < 0 < x_+ < \sqrt{b^2+4ac}/a$.

Let $\delta > 0$ and $n \geq 1$. We construct an event with probability at least $1 - \delta$ on which the support recovery property and the concentration inequality for the OLS estimate hold simultaneously. Recall that $S^\star = \mathrm{supp}(\beta^\star(f))$ is the true set of $\ell^\star = |S^\star|$ active control variables.

– Thanks to Theorem 3, with probability at least $1 - \delta/2$,

$$\hat{S}_N = S^\star. \tag{58}$$

Indeed, the conditions on N and λ in Theorem 4 are such that we can apply Theorem 3 with n and δ replaced by N and $\delta/2$, respectively.

– Thanks to Theorem 1, with probability at least $1 - \delta/2$,

$$\left|\hat{\alpha}_n^{\mathrm{ols}}(f, h_{S^\star}) - P(f)\right| \leq \sqrt{2\log(16/\delta)}\frac{\tau}{\sqrt{n}} + 58\sqrt{B^\star \ell^\star \log(16\ell^\star/\delta)\log(8/\delta)}\frac{\tau}{n}. \tag{59}$$

where for any $S \subset \{1, \ldots, m\}$, $\hat{\alpha}_n^{\mathrm{ols}}(f, h_S)$ is the OLS estimate of $P(f)$ based on the control variates h_S. Indeed, we apply Theorem 1 with h and δ replaced by $h_{S^\star}$ and $\delta/2$, respectively. The required lower bound on n is now $n \geq \max\left(18B^\star \log(8\ell^\star/\delta), 75\ell^\star \log(8/\delta)\right)$. By assumption we have $N \geq 75[\ell^\star(U_h^2/\gamma^{\star\star})]^2 \log(20\ell^\star/\delta)$. The required lower bound is already satisfied for N, and thus certainly by n.

By the union bound, the event on which (58) and (59) are satisfied simultaneously has probability at least $1-\delta$. On this event, we can, by definition of $\hat{\alpha}_n^{\mathrm{lslasso}}(f)$ and by (58), write the integration error as

$$\begin{aligned}\left|\hat{\alpha}_n^{\mathrm{lslasso}}(f) - P(f)\right| &= \left|\hat{\alpha}_n^{\mathrm{ols}}(f, h_{\hat{S}_N}) - P(f)\right| \\ &= \left|\hat{\alpha}_n^{\mathrm{ols}}(f, h_{S^\star}) - P(f)\right|.\end{aligned}$$

But the right-hand side is bounded by (59), yielding (13), as required.

References

Asuncion, A., Newman, D.: UCI machine learning repository. (2007) https://archive.ics.uci.edu/ml/index.php

Avramidis, A.N., Wilson, J.R.: A splitting scheme for control variates. Op. Res. Lett. **14**(4), 187–198 (1993)

Belloni, A., Chernozhukov, V.: Least squares after model selection in high-dimensional sparse models. Bernoulli **19**(2), 521–547 (2013)

Belomestny, D., Iosipoi, L., Moulines, E., Naumov, A., Samsonov, S.: Variance reduction for Markov chains with application to MCMC. Stat. Comput. **30**, 973–997 (2020)

Bickel, P.J., Ritov, Y., Tsybakov, A.B.: Simultaneous analysis of Lasso and Dantzig selector. Ann. Stat. **37**(4), 1705–1732 (2009)

Boucheron, S., Lugosi, G., Massart, P.: Conc. Inequal. Oxford University Press, Oxford (2013)

Brooks, S.P., Catchpole, E.A., Morgan, B.J.T.: Bayesian animal survival estimation. Stat. Sci. **15**(4), 357–376 (2000)

Brosse, N., Durmus, A., Meyn, S., Moulines, É., Radhakrishnan, A.: Diffusion approximations and control variates for MCMC. (2018) arXiv preprint arXiv:1808.01665

Caflisch, R.E.: Monte Carlo and quasi-Monte Carlo methods. Acta Numer. **7**, 1–49 (1998)

Davis, R.A., do Rêgo Sousa, T., Klüppelberg, C.: Indirect inference for time series using the empirical characteristic function and control variates. J. Time Ser. Anal. (2019)

Glasserman, P.: Monte Carlo Methods in Financial Engineering, vol. 53. Springer, Berlin (2013)

Glynn, P.W., Szechtman, R.: Some new perspectives on the method of control variates. In: Monte Carlo and Quasi-Monte Carlo Methods 2000, Springer, pp 27–49 (2002)

Gobet, E., Labart, C.: Solving BSDE with adaptive control variate. SIAM J. Numer. Anal. **48**(1), 257–277 (2010)

Gorman, R.P., Sejnowski, T.J.: Analysis of hidden units in a layered network trained to classify sonar targets. Neural Netw. **1**(1), 75–89 (1988)

Gower, R., Le Roux, N., Bach, F.: Tracking the gradients using the hessian: A new look at variance reducing stochastic methods. In: International Conference on Artificial Intelligence and Statistics, PMLR, pp 707–715 (2018)

Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)

Hsu, D., Kakade, S.M., Zhang, T.: Random design analysis of ridge regression. In: Conference on learning theory, JMLR Workshop and Conference Proceedings, pp 9–1 (2012)

Hsu, D., Kakade, S.M., Zhang, T.: Random design analysis of ridge regression. Found. Comput. Math. **14**(3), 569–600 (2014)

Javanmard, A., Montanari, A.: Debiasing the lasso: optimal sample size for Gaussian designs. Ann. Stat. **46**(6A), 2593–2622 (2018)

Jie, T., Abbeel, P.: On a connection between importance sampling and the likelihood ratio policy gradient. In: Advances in Neural Information Processing Systems, pp 1000–1008 (2010)

Jin, C., Netrapalli, P., Ge, R., Kakade, S.M., Jordan, M.: A short note on concentration inequalities for random vectors with subGaussian norm. arXiv preprint arXiv:1902.03736 (2019)

Lebreton, J.D., Burnham, K.P., Clobert, J., Anderson, D.R.: Modeling survival and testing biological hypotheses using marked animals: a unified approach with case studies. Ecol Monogr **62**(1), 67–118 (1992)

Liu, H., Feng, Y., Mao, Y., Zhou, D., Peng, J., Liu, Q.: Action-dependent control variates for policy optimization via stein identity. In: ICLR 2018 Conference (2018)

Marzolin, G.: Polygynie du Cincle plongeur (Cinclus cinclus) dans les côtes de Lorraine. Oiseau et la Revue Francaise d'Ornithologie **58**(4), 277–286 (1988)

Newey, W.K.: Convergence rates and asymptotic normality for series estimators. J. Econom. **79**(1), 147–168 (1997)

Nott, D.J., Drovandi, C.C., Mengersen, K., Evans, M., et al.: Approximation of Bayesian predictive p-values with regression ABC. Bayesian Anal. **13**(1), 59–83 (2018)

Oates, C.J., Girolami, M., Chopin, N.: Control functionals for Monte Carlo integration. J. R. Stat. Soc. Ser. B **79**(3), 695–718 (2017). (Statistical Methodology)(Statistical Methodology)(Statistical Methodology)**(Statistical Methodology)**

Owen, A., Zhou, Y.: Safe and effective importance sampling. J. Am. Stat. As. **95**(449), 135–143 (2000)

Owen, A.B.: Monte Carlo Theory, Methods and Examples (2013)

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)

Portier, F., Delyon, B.: Asymptotic optimality of adaptive importance sampling. Adv. Neural Inf. Process. Syst. **31**, 3134–3144 (2018)
Portier, F., Segers, J.: Monte Carlo integration with a growing number of control variates. J. Appl. Probab. **56**, 1168–1186 (2019)
Ranganath, R., Gerrish, S., Blei, D.: Black box variational inference. In: Artificial Intelligence and Statistics, pp 814–822 (2014)
Rudin, W.: Real and Complex Analysis. Tata McGraw-Hill Education, New York (2006)
South, L.F., Oates, C.J., Mira, A., Drovandi, C.: Regularised zero-variance control variates for high-dimensional variance reduction. arXiv preprint arXiv:1811.05073 (2018)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser.B **58**(1), 267–288 (1996). **Methodological**
Tibshirani, R., Wainwright, M., Hastie, T.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC, Boca Raton (2015)
Tropp, J.A.: An introduction to matrix concentration inequalities. Foundations and Trends® in Machine Learning 8(1-2):1–230, (2015) arXiv:1501.01571
van de Geer, S.A., Bühlmann, P.: On the conditions used to prove oracle results for the Lasso. Electr. J. Stat. **3**, 1360–1392 (2009)
Wang, C., Chen, X., Smola, A.J., Xing, E.P.: Variance reduction for stochastic gradient optimization. In: Advances in Neural Information Processing Systems, pp 181–189 (2013)
Zhang, A., Brown, L.D., Cai, T.T.: Semi-supervised inference: general theory and estimation of means. Ann. Stat. **47**(5), 2538–2566 (2019)