A data-dependent weighted LASSO under Poisson noise

Xin J. Hunt,
SAS Institute Inc., Cary, NC USA

Patricia Reynaud-Bouret
University of Côte d'Azur, CNRS, LJAD, Nice, France

Vincent Rivoirard
University of Paris-Dauphine, Paris, France

Laure Sansonnet
INRA - AgroParisTech, Paris, France

Rebecca Willett*
University of Wisconsin-Madison, Madison, WI, USA

September 2, 2018

Abstract

Sparse linear inverse problems appear in a variety of settings, but often the noise contaminating observations cannot accurately be described as bounded by or arising from a Gaussian distribution. Poisson observations in particular are a characteristic feature of several real-world applications. Previous work on sparse Poisson inverse problems encountered several limiting technical hurdles. This paper describes a novel alternative analysis approach for sparse Poisson inverse problems that (a) sidesteps the technical challenges present in previous work, (b) admits estimators that can readily be computed using off-the-shelf LASSO algorithms, and (c) hints at a general framework for broad classes of noise in sparse linear inverse problems. At the heart of this new approach lies a weighted LASSO estimator for which *data-dependent* weights are based on Poisson concentration inequalities. Unlike previous analyses of the weighted LASSO, the proposed analysis depends on conditions which can be checked or shown to hold in general settings with high probability.

Keywords: Weighted LASSO, Poisson Noise, Compressed Sensing, Genetic Motifs, Photon-Limited Imaging
2000 Math Subject Classification: 60E15, 62G05, 62G08, 94A12

1 Introduction

Poisson noise arises in a wide variety of applications and settings, including PET, SPECT, and pediatric or spectral CT [1, 2, 3] in medical imaging, x-ray astronomy [4, 5, 6], genomics [7], network packet analysis [8, 9], crime rate analysis [10], and social media analysis [11]. In these and other settings, observations are characterized by discrete counts of events (*e.g.,* photons

*This paper was presented in part at the 2016 IEEE Statistical Signal Processing Workshop.

hitting a detector or packets arriving at a network router), and our task is to infer the underlying signal or system even when the number of observed events is very small. Methods for solving Poisson inverse problems have been studied using a variety of mathematical tools, with recent efforts focused on leveraging signal sparsity [12, 13, 14, 15, 2, 6, 16, 17].

Unfortunately, the development of risk bounds for sparse Poisson inverse problems presents some significant technical challenges. Methods that rely on the negative Poisson log-likelihood to measure how well an estimate fits observed data perform well in practice but are challenging to analyze. For example, the analysis framework considered in [12, 13, 14] builds upon a coding-theoretic bound which is difficult to adapt to many of the computationally tractable sparsity regularizers used in the Least Absolute Shrinkage and Selection Operator (LASSO) [18] or Compressed Sensing (CS) [19, 20]; those analyses have been based on impractical ℓ_0 sparsity regularizers. In contrast, the standard LASSO analysis framework easily handles a variety of regularization methods and has been generalized in several directions [21, 22, 23, 18, 24, 25]. However, it does not account for Poisson noise, which is heterogeneous and dependent on the unknown signal to be estimated.

This paper presents an alternative approach that sidesteps these challenges. We describe a novel weighted LASSO estimator, where the data-dependent weights used in the regularizer are based on Poisson concentration inequalities and control for the ill-posedness of the inverse problem and heteroscedastic noise simultaneously. We establish oracle inequalities and recovery error bounds for general settings, and then explore the nuances of our approach within two specific sparse Poisson inverse problems arising in genomics and imaging.

1.1 Problem formulation

We observe a potentially random matrix $A = (a_{k,l})_{k,l} \in \mathbb{R}_+^{n\times p}$ and conditionally on A, we observe

$$Y \sim \mathcal{P}(Ax^*) \tag{1.1}$$

where $Y \in \mathbb{R}_+^n$, $x^* \in \mathbb{R}_+^p$, and where x^* is sparse or compressible. The notation $\mathcal{P}$ denotes the Poisson distribution, so that, conditioned on A and x^*, we have the likelihood

$$p(Y_k|Ax^*) = e^{-(Ax^*)_k}[(Ax^*)_k]^{Y_k}/Y_k!, \qquad k = 1, \dots, n.$$

Conditioned on Ax^*, the elements of Y are independent. The aim is to recover x^*, the true signal of interest. The matrix A corresponds to a sensing matrix or operator which linearly projects x^* into another space before we collect Poisson observations. Often we will have $n < p$, but this inverse problem can still be challenging if $n \geq p$ depending on the signal-to-noise ratio or the condition of the operator A.

Because elements of A are nonnegative, we cannot rely on the standard assumption that $A^\top A$ is "close to" an identity matrix. However, in many settings there is a proxy operator, denoted $\widetilde{A}$, which is amenable to sparse inverse problems and is a simple linear transformation of the original operator A. A complementary linear transformation may then be applied to Y to generate proxy observations $\widetilde{Y}$, and we use $\widetilde{A}$ and $\widetilde{Y}$ in the estimators defined below. In general, the linear transformations are problem-dependent, and may depend on A. However, the linear transformations do not depend on Y. Thus, once A is fixed, the transformations are deterministic. The transformations should be chosen to ensure our main assumptions (presented in Section 2) are satisfied.

Specifically, $\widetilde{A}$ and $\widetilde{Y}$ are often chosen so that the corresponding Gram matrix $\widetilde{G} = \widetilde{A}^\top \widetilde{A}$ is as homogeneous as possible, and that $\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)$ is as small as possible. We refer the reader to discussions in Sections 1.3 and 3. We provide explicit examples in Sections 4 and 5. Note that other preconditioning transformations have been proposed in the literature. See for instance [26, 27, 28] where various procedures are suggested but very different in spirit from ours.

1.2 Weighted LASSO estimator for Poisson inverse problems

The basic idea of our approach is the following. We consider two main estimation methods in this paper:

(Classical) LASSO estimator:

$$\widehat{x}^{\mathrm{LASSO}} := \operatorname*{argmin}_{x\in\mathbb{R}^p} \left\{ \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma d \|x\|_1 \right\}, \tag{1.2}$$

where $\gamma > 2$ is a constant and $d > 0$ is a data-dependent scalar to be defined later.

Weighted LASSO estimator:

$$\widehat{x}^{\mathrm{WL}} := \operatorname*{argmin}_{x\in\mathbb{R}^p} \left\{ \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=1}^{p} d_k |x_k| \right\}, \tag{1.3}$$

where $\gamma > 2$ is a constant and the d_k's are positive and data-dependent; they will be defined later. Note that the estimator in (1.3) can equivalently be written as

$$\widehat{z} = \operatorname*{argmin}_{z\in\mathbb{R}^p} \left\{ \|\widetilde{Y} - \widetilde{A}D^{-1}z\|_2^2 + \gamma \|z\|_1 \right\}, \tag{1.4a}$$

$$\widehat{x}^{\mathrm{WL}} = D^{-1}\widehat{z}, \tag{1.4b}$$

where D is a diagonal matrix with the k^{th} diagonal element equal to d_k. Note that the optimization problem in (1.4a) can be solved efficiently using off-the-shelf LASSO solvers. Since z and $D^{-1}z$ will always have the same support, this formulation suggests that the weighted LASSO estimator in (1.3) is essentially a data-dependent reweighing of the columns of $\widetilde{A}$.

A weighted LASSO estimator similar to (1.3) has been proposed and analyzed in past work, notably [29, 21, 30], where the weights are considered fixed and arbitrary. The analysis in [29], however, does not extend to signal-dependent noise (as we have in Poisson noise settings). In addition, risk bounds in that work hinge on a certain "*weighted* irrepresentable condition" on the sensing or design matrix $\widetilde{A}$ which cannot be verified or guaranteed for the data-dependent weights we consider, even when $\widetilde{A}$ is known to satisfy criteria such as the Restricted Eigenvalue condition [21] or Restricted Isometry Property [31]. An analysis of the weighted LASSO estimator using standard LASSO bounding techniques (*cf.* [30, 32]) yields looser bounds than those presented below.

If x^* has support S^* of size $s := |S^*|$ and if we choose weights $d_1, \ldots, d_p$ satisfying

$$|(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k| \le d_k \qquad \text{for } k = 1, \ldots, p, \tag{1.5}$$

then, over an appropriate range of values of s, the following risk bounds hold for the LASSO and weighted LASSO estimates under conditions of Proposition 2.1:

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_2^2 \leq \frac{\rho_\gamma^2}{\eta} \sum_{k \in S^*} d_k^2, \tag{1.6a}$$

$$\|\widehat{x}^{\mathrm{LASSO}} - x^*\|_2^2 \leq \frac{\rho_\gamma^2}{\eta} s d^2, \tag{1.6b}$$

where ρ_γ only depends on γ and η is a parameter associated with the restricted eigenvalue condition of the sensing matrix $\tilde{A}$ (see Proposition 2.1). Note that the condition in (1.5) is similar to the constraint in the Dantzig selector [33]. The bounds in (1.6) highlight that if we did not have practical constraints such as the fact that the d_k's should only depend on the data, one could take $d_k = |(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k|$ for the weighted LASSO, whereas for the LASSO, one could only take $d = \max_k |(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k|$, which only leads to worse bounds. Note that if the optimization problems in Equation (1.3) and Equation (1.4a) have more than one minimizer, the proposed bounds in Inequality (1.6) hold for every minimizer.

Furthermore, we consider an oracle estimator that consists of least squares estimation on the true support S^*,

$$\widehat{x}^{\mathrm{OLS}} := I_{S^*}(\widetilde{A}_{S^*})^{\#}\widetilde{Y},$$

and show that

$$\sum_{k \in S^*} (\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k^2 \lesssim \|\widehat{x}^{\mathrm{OLS}} - x^*\|_2^2 \lesssim \sum_{k \in S^*} (\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k^2.$$

If each d_k in the weighted LASSO estimator is close to $|(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k|$, then the bounds on the weighted LASSO are close to those of the oracle least squares estimator.

In practice the weights can only depend on the observed data. We show two examples, a Bernoulli sensing matrix and random convolution (see Sections 4 and 5), in which we can compute weights from the data (by using Poisson concentration inequalities) such that (1.5) holds with high probability *and* those weights are small enough to ensure risk bounds that have a better convergence rate than LASSO estimates.

1.3 The role of the weights

Our approach, where the weights in our regularizer are random variables, is similar to [34, 35, 36, 37]. In some sense, the weights play the same role as the thresholds in the estimation procedure proposed in [38, 39, 40, 41, 7]. The role of the weights are twofold:

- control of the random fluctuations of $\widetilde{A}^\top\widetilde{Y}$ around its mean, and
- compensate for the ill-posedness due to $\widetilde{A}$. Note that ill-posedness is strengthened by the heteroscedasticity of the Poisson noise.

To better understand the role of the weights, let us look at a toy example where A is a diagonal matrix with decreasing eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ to which we add heteroscedastic noise. Here one could rephrase the direct problem as

$$y_k = \lambda_k x_k^* + \epsilon_k, \tag{1.7}$$

where ϵ_k has zero mean and standard deviation σ_k. This toy example is often derived by diagonalizing some inverse problem via the singular value decomposition of the matrix A[1]. The assumptions of Poisson noise and that x^* is sparse do not generally hold if we diagonalize the problem in (1.1), so the model in (1.7) should not be seen as a sketch of what can be done in general but more as an illustration to understand the main tools that are used in the sequel.

First of all, note that if we know the support of x^*, the least-square estimator, $\hat{x}^{LS}$, is trivial here and amounts to $(\hat{x}^{LS})_k = y_k/\lambda_k$ if k is in S^* and 0 anywhere else. It is then easy to see that

$$\mathbb{E}(\|\hat{x}^{LS} - x^*\|_2^2) = \sum_{k \in S^*} \frac{\sigma_k^2}{\lambda_k^2}.$$

This is our benchmark[2], our oracle in some sense, since it cannot be computed without knowing the support of the true signal.

The weighted LASSO method needs two main ingredients: the linear transformation $(\tilde{A}, \tilde{Y})$ and the weights satisfying (1.5). Let us first consider the ramifications of setting $\widetilde{Y} = Y$, $\widetilde{A} = A$. The quantity η appearing in (1.6a) is then connected to the smallest eigenvalue of A, and it is easy to show[3] that $\eta = \lambda_p^4$. On the other hand, by (1.5), d_k should be an upper bound on $\lambda_k \epsilon_k$. Heuristically, d_k should therefore be of the order of $\lambda_k \sigma_k$ and the bound (1.6a) is then on the order of $\sum_{k \in S^*} \frac{\lambda_k^2 \sigma_k^2}{\lambda_p^4}$. In particular, even if the true support S^* coincides with indices where the λ_k's are large, we still see our rate controlled by the potentially large factor of λ_p^{-4}.

On the other hand, the classical inverse problem choice $\widetilde{Y} = A^{-1}Y$, $\widetilde{A} = A^{-1}A = I_p$ gives that $\eta = 1$ and that d_k should be of the order of σ_k/λ_k at least[4]. Therefore the upper bound (1.6a) is then of the order of $\sum_{k \in S^*} \frac{\sigma_k^2}{\lambda_k^2}$, that is, we reach the benchmark risk of the least-square estimator. For the interesting case where the λ_k's for $k \in S^*$'s are much larger than λ_p, by using the weighted LASSO procedure, we only pay for ill-posedness in the support of x^*, without even knowing this support. Note also that in this set-up, if one wants to choose a constant weight d, then $d \simeq \max_k (\sigma_k/\lambda_k)$ and one again pays for global ill-posedness and not just ill-posedness in the support of x^*.

This toy example shows us three things:

(i) The d_k's are indeed balancing both ill-posedness and heteroscedasticity of the problem.

(ii) The choice of the mappings from A to $\widetilde{A}$ and Y to $\widetilde{Y}$ impacts the rates.

(iii) The non-constant d_k's allow for "adaptivity" with respect to the local ill-posedness of the problem, in terms of the support of x^*.

[1] Specifically, if the SVD of A is $U\Lambda V^\top$, then observations of the form $y = Ax^* + \epsilon$ can be equivalently expressed as $U^\top y = \Lambda(V^\top x^*) + U^\top \epsilon$, yielding measurements of the form (1.7).

[2] In addition, note that in a Gaussian framework, the least-square estimator is also the MLE and reaches Cramer-Rao bound. So in a certain sense, asymptotically it is the smallest risk we could hope for.

[3] In this diagonal case, it is easy to see that Assumption RE holds with $\kappa_2 = \lambda_p$ and $\kappa_1 = 0$, which leads to $\epsilon = \kappa_2$ in Proposition 2.1

[4] In practice, because there will be randomness to take into account, guaranteeing (1.5) for all k with high probability, will lead to an extra $\log(p)$ factor here, that may be thought as the price to pay for adaptivity with respect to unknown support S^* (see in particular the two main examples).

Of course, this example is just a toy example and many simplifications occur due to the diagonalization effect; however, the same phenomena appear in the much more intricate examples (Bernoulli and Convolution) of Sections 4 and 5. To deal with these settings, we need to choose $\widetilde{A}$ so that the corresponding Gram matrix $\widetilde{G} = \widetilde{A}^\top \widetilde{A}$ is as homogeneous as possible (meaning that there is no great discrepancy between maximal and minimal eigenvalues on restricted sets typically and informally that it looks as much as possible as the identity matrix up to a multiplicative constant) and choose $\widetilde{Y}$ such that $\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)$ is as small as possible, which will make the d_k's as small as possible and therefore giving the best possible rates, that could not be achieved using a single constant weight d. This choice in particular will enable us to get rates consistent with the minimax rates derived in [12] in a slightly different framework.

1.4 Organization of the paper

Section 2 describes general oracle inequalities, recovery rate guarantees, and support recovery bounds for the three estimators described above, given weights which satisfy (1.5). We then describe a general framework for finding such weights using the observed data in Section 3. We next describe exact weights and resulting risk bounds for two specific Poisson inverse problems: (a) Poisson compressed sensing using a Bernoulli sensing matrix, which models certain optical imaging systems such as [42], and (b) a ill-posed Poisson deconvolution problem arising in genetic motif analysis, building upon the formulation described in [7]. We conclude with simulation-based verification of our derived rates.

1.5 Notation

To provide readable results, we use the following notation in the sequel: $a \lesssim_\gamma b$ if there exists a positive constant c_γ only depending on γ such that $a \leq c_\gamma b$. Similarly, $a \gtrsim_\gamma b$ means $b \lesssim_\gamma a$ and $a \simeq_\gamma b$ means both $a \lesssim_\gamma b$ and $a \gtrsim_\gamma b$. If there is no index γ, it just means that the constants are absolute. We use $a = o\left(b\right)$ if the ratio a/b goes to 0 asymptotically, and $a = \omega\left(b\right)$ if the ratio a/b goes to $+\infty$ asymptotically. In the proofs, the notation $\square$ represents an absolute constant that may change from line to line. $\mathbb{1}_{m\times n}$ represents an all ones matrix of size $m \times n$, whereas $\mathbb{1}_m$ is a shorthand for $\mathbb{1}_{m\times 1}$. For any vector $z \in \mathbb{R}^p$ and $S \subseteq \{1, \ldots, p\}$, let $z_S \in \mathbb{R}^p$ be defined via $(z_S)_i = \begin{cases} z_i, & i \in S \\ 0, & \text{otherwise} \end{cases}$.

2 Theoretical performance bounds for the weighted LASSO

In this section, we establish recovery error bounds for the proposed weighted LASSO estimator. The underlying proof techniques closely follow those described in [32, 43] and elsewhere, but have been adapted to account for the weighted-ℓ_1 regularizer. Without this adaptation, directly applying the theory of [32] to the weighted LASSO estimator yields rates that are equivalent to those of the standard LASSO estimator, modulo a constant factor. As shown below, our modified analysis yields tighter bounds that better reflects the role of the weights, as discussed in detail in the examples in the following sections. We state our bounds in this section and the associated proofs in the appendix for completeness and clarity. The bounds in this section do not depend on the noise distribution and can be used regardless of the underlying noise. They rely upon the following two main assumptions, both of which are proved to be met with high probability in two key examples in the next sections.

The first assumption is known as the *Restricted Eigenvalue Condition* (see [32, 44, 45]):

ASSUMPTION RE(κ_1, κ_2) There exist $\kappa_2, \kappa_1 > 0$ such that

$$\|\widetilde{A}x\|_2 \geq \kappa_2 \|x\|_2 - \kappa_1 \|x\|_1 \qquad \forall\, x \in \mathbb{R}^p. \tag{2.1}$$

This condition is weaker than the verifiable condition in [30]. Our other key assumption dictates the necessary relationship between the weights used to regularize the estimates $\widehat{x}^{\mathrm{WL}}$ and $\widehat{x}^{\mathrm{LASSO}}$.

ASSUMPTION WEIGHTS($\{d_k\}_k$) For $k = 1, \ldots, p$,

$$|(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k| \leq d_k. \tag{2.2}$$

In the sequel, we also use the following definitions:

$$d_{\max} := \max_{k \in \{1,\ldots,p\}} d_k, \qquad d_{\min} := \min_{k \in \{1,\ldots,p\}} d_k, \quad \text{and} \quad \rho_\gamma := \gamma \frac{\gamma+2}{\gamma-2}. \tag{2.3}$$

Further recall that D is a diagonal matrix with the k^{th} diagonal element equal to d_k. Because D is diagonal, note that for any vector $z \in \mathbb{R}^p$ and any set $S \subseteq \{1, \ldots, p\}$, $Dz_S = (Dz)_S$.

Proposition 2.1. *Fix $\varepsilon > 0$. If $\gamma > 2$ and Assumptions Weights($\{d_k\}_k$) and RE(κ_1, κ_2) are satisfied, then there exists a universal constant $c > 0$ such that for any set $S \subseteq \{1, \ldots, p\}$ for which*

$$\|d_S\|_2 \leq d_{\min} \frac{\kappa_2 - \varepsilon}{\kappa_1 \rho_\gamma}, \tag{2.4}$$

the weighted LASSO estimator satisfies

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_2 \leq c \left(\frac{\rho_\gamma}{\varepsilon^2} \|d_S\|_2 + \sqrt{\frac{\rho_\gamma}{\varepsilon^2}} \|Dx^*_{S^c}\|_1^{1/2} + \frac{\rho_\gamma \kappa_1}{\varepsilon d_{\min}} \|Dx^*_{S^c}\|_1 \right). \tag{2.5}$$

Furthermore, for any set $S \subseteq \{1, \ldots, p\}$ with $s = |S|$ satisfying

$$\sqrt{s} \leq \frac{\kappa_2 - \varepsilon}{\kappa_1 \rho_\gamma}, \tag{2.6}$$

for $\varepsilon > 0$, the LASSO estimator satisfies

$$\|x^* - \widehat{x}^{\mathrm{LASSO}}\|_2 \leq c \left(\frac{\rho_\gamma}{\varepsilon^2} d\sqrt{s} + \sqrt{\frac{\rho_\gamma d}{\varepsilon^2}} \|x^*_{S^c}\|_1^{1/2} + \frac{\rho_\gamma \kappa_1}{\varepsilon} \|x^*_{S^c}\|_1 \right). \tag{2.7}$$

If the weights d_k are constant, Inequality (2.4) is equivalent to Inequality (2.6), which is natural as $\widehat{x}^{\mathrm{WL}}$ is equivalent to $\widehat{x}^{\mathrm{LASSO}}$ when $d_k = d, k = 1, \ldots, p$. Note that Inequality (2.5) can be expressed simply in the case where the vector x^* is s-sparse with support S^* where $s = |S^*|$ and S^* satisfies (2.4):

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_2 \leq c \frac{\rho_\gamma}{\varepsilon^2} \|d_{S^*}\|_2. \tag{2.8}$$

This result clearly shows the importance of having weights as small as possible but large enough so that Assumption (Weights) is satisfied and not too heterogeneous so that (2.4) is true. This trade-off for weights choice will be illustrated in Sections 4 and 5. Note that (2.4) is equivalent to

$$\sqrt{s} \leq \frac{\|d_S\|_2}{d_{\min}} \leq \frac{\kappa_2 - \varepsilon}{\kappa_1 \rho_\gamma}$$

with $s = |S|$, so that this condition is both a sparsity condition and a limit on the heterogeneity of the weights.

3 Choosing data-dependent weights

In general, choosing d_k's to ensure that Assumption Weights($\{d_k\}_k$) is satisfied is highly problem-dependent, and we give two explicit examples in the following two sections. In this section we present the general strategy we adopt for choosing the weights. The weights d_k are ideally chosen so that for all k

$$|(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k| \leq d_k. \tag{3.1}$$

We describe a data-dependent strategy for choosing weights such that this condition holds with high probability. The modifications $\widetilde{Y}$ and $\widetilde{A}$ of Y and A that we have in mind are linear, therefore one can generally rewrite for each k,

$$(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k = R_k^\top(Y - Ax^*) + r_k(A, x^*),$$

for some vector $R_k \in \mathbb{R}^n$ which depends on k and A, and for some residual term $r_k(A, x^*)$, also depending on k and A. The transformations are chosen such that d_k is small. Recall the observation model in Equation (1.1). With the above decomposition, the first term $R_k^\top(Y - Ax^*)$ is naturally of null conditional expectation given A and therefore of zero mean. The $\widetilde{Y}$ are usually chosen such that $r_k(A, x^*)$ is also of zero mean which globally guarantees that $\mathbb{E}[(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k] = 0$.

In the two following examples, the term $r_k(A, x^*)$ is either mainly negligible with respect to $R_k^\top(Y - Ax^*)$ (Bernoulli case, Section 4) or even identically zero (convolution case, Section 5). Therefore the weights are mainly given by concentration formulas on quantities of the form $R^\top(Y - Ax^*)$ as given by the following Lemma.

Lemma 3.1. *With observation model from Equation (1.1), for all vectors $R = (R_k)_{k=1,\dots,n} \in \mathbb{R}^n$, eventually depending on A, let $\bar{R}_2 := (R_k^2)_{k=1,\dots,n}$. Then the following inequality holds for all $\theta > 0$,*

$$\mathbb{P}\left(R^\top Y \geq R^\top Ax^* + \sqrt{2v\theta} + \frac{b\theta}{3} \,\middle|\, A\right) \leq e^{-\theta}, \tag{3.2}$$

with

$$v = \bar{R}_2^\top \mathbb{E}(Y|A) = \bar{R}_2^\top Ax^*$$

and

$$b = \|R\|_\infty.$$

Moreover

$$\mathbb{P}\left(|R^\top Y - R^\top Ax^*| \geq \sqrt{2v\theta} + \frac{b\theta}{3} \,\middle|\, A\right) \leq 2e^{-\theta}, \tag{3.3}$$

$$\mathbb{P}\left(v \geq \left(\sqrt{\frac{b^2\theta}{2}} + \sqrt{\frac{5b^2\theta}{6} + \bar{R}_2^\top Y}\right)^2 \,\Big|\, A\right) \leq e^{-\theta}, \tag{3.4}$$

and

$$\mathbb{P}\left(|R^\top Y - R^\top Ax^*| \geq \left(\sqrt{\frac{b^2\theta}{2}} + \sqrt{\frac{5b^2\theta}{6} + \bar{R}_2^\top Y}\right)\sqrt{2\theta} + \frac{b\theta}{3} \,\Big|\, A\right) \leq 3e^{-\theta}. \tag{3.5}$$

Inequalities (3.2) and (3.3) give the main order of magnitude for $R^\top(Y - Ax^*)$ with high probability but are not sufficient for our purpose since v still depends on the unknown x^*. That is why Inequality (3.4) provides an estimated upper-bound for v with high probability. Inequality (3.5) is therefore our main ingredient for giving observable d_k's that satisfy Assumption Weights($\{d_k\}_k$). Note that, depending on A, one may also find a more particular way to define those weights, in particular constant ones. This is illustrated in the two examples in Sections 4 and 5.

3.1 Overview of general approach

It is worth noting that the core results of Section 2 do not use any probabilistic arguments and therefore do not rely at all on Poisson noise assumptions. The Poisson noise model is only used to derive data-dependent weights that satisfy the necessary assumptions with high probability under the assumed observation model. To extend our framework to new observation or noise models, we would simply need to complete the following (interdependent) tasks:

1. Determine a mapping from A to $\widetilde{A}$ which ensures $\widetilde{A}$ satisfies Assumption RE.
2. Determine a mapping from Y to $\widetilde{Y}$ so that $\mathbb{E}[\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)] = 0$.
3. Use concentration inequalities based on the assumed noise model to derive data-dependent weights which satisfy Assumption Weights.

Once these tasks are complete, the results of Section 2 can be immediately applied to compute recovery error rates. Thus the proposed weighted LASSO framework has potential for a variety of settings and noise distributions.

4 Case Study: Photon-limited compressive imaging

A widely-studied compressed sensing measurement matrix is the Bernoulli or Rademacher ensemble, in which each element of A is drawn iid from a Bernoulli(q) distribution for some $q \in (0, 1)$. (Typically, $q = 1/2$.) In fact, the celebrated Rice single-pixel camera [42] uses exactly this model to position the micromirror array for each projective measurement. This sensing matrix model has also been studied in previous work on Poisson compressed sensing (*cf.* [12, 13]). In this section, we consider our proposed weighted LASSO estimator for this sensing matrix. Because our focus is a comparison of the classical and weighted LASSO estimators, we focus here on s-sparse x^*, which is consistent with previous theoretical analyses of this problem. Note, however, that our results extend trivially to the non-sparse setting.

4.1 Rescaling and recentering

Our first task is to define the surrogate design matrix $\widetilde{A}$ and surrogate observations $\widetilde{Y}$. In this set-up, one can easily see that the matrix

$$\widetilde{A} = \frac{A}{\sqrt{nq(1-q)}} - \frac{q\mathbb{1}_{n\times 1}\mathbb{1}_{p\times 1}^\top}{\sqrt{nq(1-q)}} \tag{4.1}$$

is a scaled and shifted version of the original A and satisfies $\mathbb{E}(\widetilde{A}^\top \widetilde{A}) = I_p$ (see Appendix C.1), which will help us to ensure that Assumption RE holds. To make d_k as small as possible while still satisfying Assumption Weights($\{d_k\}_k$), we would like to have $\mathbb{E}(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)) = 0$, as stated previously. Computations given in Appendix C.1 show that it is sufficient to take

$$\widetilde{Y} = \frac{1}{(n-1)\sqrt{nq(1-q)}}(nY - \sum_{\ell=1}^{n} Y_\ell \mathbb{1}_{n\times 1}). \tag{4.2}$$

4.2 Assumption RE holds with high probability.

Proposition 4.1. *There exist positive absolute constants c' and c'' such that with probability larger than $1 - c'\exp(-c''n)$, Assumption RE(κ_1, κ_2) holds with*

$$\kappa_1 = \frac{c}{q(1-q)}\sqrt{\frac{\log p}{n}} \quad \text{and} \quad \kappa_2 = \frac{1}{4},$$

where c is an absolute positive constant.

Since q and n effectively determine the signal-to-noise ratio, it is generally more interesting to understand the performance guarantee when q is small. In the sequel, we focus on the interesting case when q is small and tends to 0 as n and p grow. Therefore we assume that

$$0 < q \leq \frac{1}{2} \leq 1 - q < 1. \tag{4.3}$$

In addition to q being small, the orders of magnitude that we have derived only hold for

$$q \gtrsim \sqrt{\frac{\log(p)}{n}}, \tag{4.4}$$

which implies in particular that

$$\log(p) \lesssim n. \tag{4.5}$$

In particular, q can still tend to 0 with n and p but cannot be too small, as long as $\log(p) = o(n)$.

4.3 Choice of the weights

We now discuss the rates obtained by applying Proposition 2.1 for constant and non-constant weights that are found thanks to the machinery described in Section 3.

4.3.1 Definition of constant weight and rates for the estimate $\widehat{x}^{\mathrm{LASSO}}$

For all $k = 1, \dots, p$, define the vector $V_k \in \mathbb{R}^n$ so that the ℓ^{th} element is

$$V_{k,\ell} := \left(\frac{n a_{\ell,k} - \sum_{\ell'=1}^n a_{\ell',k}}{n(n-1)q(1-q)} \right)^2 \tag{4.6}$$

and let $A_k \in \mathbb{R}^n$ denote the k^{th} column of A.

Let

$$W = \max_{u,k \in \{1,\dots,p\}} \langle A_u, V_k \rangle$$

and

$$\hat{N} = \frac{1}{nq - \sqrt{6nq(1-q)\log(p)} - \max(q, 1-q)\log(p)} \left(\sqrt{\frac{3\log(p)}{2}} + \sqrt{\frac{5\log(p)}{2} + \sum_{\ell=1}^n Y_\ell} \right)^2$$

be an estimator of $\|x^*\|_1$. Computations of Appendix C.3.1 show that $\hat{N}$ is a sharp (observable) upper bound of $\|x^*\|_1$. One can then choose a constant weight defined by

$$d = \sqrt{\hat{N} 6W \log(p)} + \frac{\log(p)}{(n-1)q(1-q)} + c\left(\frac{3\log(p)}{n} + \frac{9\max(q^2, (1-q)^2)}{n^2 q(1-q)} \log(p)^2 \right) \hat{N}, \tag{4.7}$$

where c is an absolute constant (see the proof of Proposition C.1, $c = 126$ works if $n \geq 20$). One can prove that this satisfies Assumption Weights(d) except on an event of probability of order $1/p$ as long as $p \geq 2$ (see Proposition C.1).

If we want to give an explicit rate for the corresponding classical LASSO estimator, we need to find the order of magnitude of d. As shown in Proposition C.2, in the range (4.4), we have

$$d \simeq \sqrt{\frac{\log(p)\|x^*\|_1}{nq}} + \frac{\log(p)\|x^*\|_1}{n} + \frac{\log(p)}{nq}.$$

We now apply Proposition 2.1 with $\varepsilon = 1/8$ and a fixed $\gamma > 2$. It is easy to see that if the size of the true support s satisfies

$$s = o\left(\frac{nq^2}{\log p} \right), \tag{4.8}$$

then (2.6) holds. Hence, all assumptions of Proposition 2.1 are satisfied and we finally have the following proposition:

Proposition 4.2. *Assume that (4.3) and (4.4) are satisfied. For the constant weight defined in Equation (4.7), if (4.8) holds, then except on an event of probability of order $1/p + e^{-c"n}$ (where c" is introduced in Proposition 4.1),*

$$\|\widehat{x}^{\mathrm{LASSO}} - x^*\|_2^2 \lesssim_\gamma \frac{\log p}{n} \left(\frac{\|x^*\|_1 s}{q} + \frac{\|x^*\|_1^2 s \log p}{n} + \frac{s \log p}{nq^2} \right). \tag{4.9}$$

In particular, note that under assumptions of the previous proposition and in the range

$$1 \lesssim \|x^*\|_1 \lesssim n/\log(p), \tag{4.10}$$

the first term of (4.9) dominates and we have $\|\widehat{x}^{\mathrm{LASSO}} - x^*\|_2^2 \lesssim_\gamma \frac{s\|x^*\|_1 \log p}{nq}$.

4.3.2 Definition of non-constant weights and rates for the estimate $\widehat{x}^{\mathrm{WL}}$

One can choose the non-constant weights defined by

$$d_k = \sqrt{6\log(p)}\left(\sqrt{\frac{3\log(p)}{2(n-1)^2q^2(1-q)^2}} + \sqrt{\frac{5\log(p)}{2(n-1)^2q^2(1-q)^2} + \langle V_k, Y\rangle}\right) \\ + \frac{\log(p)}{(n-1)q(1-q)} + c\left(\frac{3\log(p)}{n} + \frac{9\max(q^2,(1-q)^2)}{n^2q(1-q)}\log(p)^2\right)\hat{N}, \qquad (4.11)$$

where V_k is defined in (4.6) and c is an absolute constant (see the proof of Proposition C.3, $c = 126$ works if $n \geq 20$). They also satisfy Assumption Weights(d) except on an event of probability of order $1/p$ as long as $p \geq 2$ (see Proposition C.3). Furthermore, as shown in Proposition C.4, in the range (4.4), we have the following order of magnitude

$$d_k \simeq \sqrt{\log(p)\left[\frac{x_k^*}{nq} + \frac{\sum_{u\neq k} x_u^*}{n}\right]} + \frac{\log(p)\|x^*\|_1}{n} + \frac{\log(p)}{nq}.$$

For S^* the support of x^*, note that

$$\frac{\|d_{S^*}\|_2}{d_{\min}} \lesssim \sqrt{s + \frac{1}{q}}.$$

Therefore (2.4) is satisfied as soon as

$$s = o\left(\frac{nq^2}{\log p}\right) \text{ and } q = \omega\left(\left(\frac{\log(p)}{n}\right)^{1/3}\right). \qquad (4.12)$$

The first part is exactly (4.8), and the second part is slightly stronger than (4.4). However, q can still tends to 0 with p and n as long as $\log(p) = o(n)$.

Under (4.12), we can now apply Proposition 2.1 as before, and obtain the following proposition:

Proposition 4.3. *Assume that (4.3) and (4.4) are satisfied. For the non-constant weight defined in Equation (4.11), if (4.12) holds, then except on an event of probability of order $1/p + e^{-c''n}$ (where c" is introduced in Proposition 4.1),*

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_2^2 \lesssim_\gamma \frac{\log p}{n}\left(\frac{\|x^*\|_1}{q} + \|x^*\|_1 s + \frac{\|x^*\|_1^2 s\log p}{n} + \frac{s\log p}{nq^2}\right). \qquad (4.13)$$

In particular, note that under assumptions of the previous proposition and in the range $1 \lesssim \|x^*\|_1 \lesssim n/\log(p)$, the first two terms of (4.13) dominate and we have

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_2^2 \lesssim_\gamma \frac{\log p\|x^*\|_1(s+1/q)}{n}.$$

Therefore, one can form the ratio of the upper bounds derived in the classical setting and the one obtained here, leading to

$$\frac{\frac{\log p}{n}\left(\frac{\|x^*\|_1}{q} + \|x^*\|_1 s\right)}{\frac{\log p}{n}\left(\|x^*\|_1 s/q\right)} \lesssim \frac{1}{s} + q,$$

which is much smaller than 1. The upper bound on the weighted LASSO estimator is therefore better in this range. The error bounds for both the LASSO and weighted LASSO are worse for smaller q (below $1/2$). As q approaches zero, we essentially get less data with which we can perform estimation. Small q may be favorable for hardware or implementation reasons, but independent of external factors like this, $q \approx 1/2$ is best. However, as the previous ratio indicates, the weighted LASSO bound is much smaller than the regular LASSO bound by a factor of $1/s + q$. Thus, while both methods struggle for smaller q, the regular LASSO struggles more.

The term $\|x^*\|_1$ in Proposition 4.3 reflects the fact that $\|x^*\|_1$ impacts the variance of the observations. Indeed, in the Gaussian setting

$$\widetilde{Y} = \widetilde{A}x^* + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n),$$

and we have the bound

$$\|\hat{x}^{LASSO} - x^*\|_2^2 \lesssim s \log p \times \sigma^2.$$

In the Poisson setting, σ^2 can be replaced by the mean of the variance of the $\widetilde{Y}_i$ and we obtain the bound

$$\begin{aligned} s \log p \times \frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}(\widetilde{Y}_i | \widetilde{A}) &\approx s \log p \times \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{Var}(Y_i | A)}{nq} && \text{(4.14a)} \\ &= s \log p \times \frac{1}{n} \sum_{i=1}^{n} \frac{\sum_j A_{ij} x_j^*}{nq} && \text{(4.14b)} \\ &\approx s \log p \times \frac{\|x^*\|_1}{n}, && \text{(4.14c)} \end{aligned}$$

where (4.14a) is a result of q being small, and (4.14c) is a result of the strong law of large numbers (SLLN), by which we have $\frac{1}{n}\sum_{i=1}^{n} A_{ij} \approx q$. The rate achieved by our weighted LASSO is, in the range $1 \lesssim \|x^*\|_1 \lesssim n/\log(p)$,

$$(s + 1/q) \log p \times \frac{\|x^*\|_1}{n},$$

which matches Eq. (4.14) except for the term $1/q$ inside the brackets. The factor $1/q$ is necessary, since the bound should certainly worsen with smaller q. However, whether the $1/q$ factor is optimal is unknown.

4.4 Comparison with the oracle least squares estimate

Let $S^* := \mathrm{supp}(x^*)$ denote the true signal support; we consider the *oracle least squares estimate* on S^*:

$$\widehat{x}^{\mathrm{OLS}} := I_{S^*} (\widetilde{A}_{S^*})^{\#} \widetilde{Y}, \tag{4.15}$$

where $\widetilde{A}_{S^*} \in \mathbb{R}^{n \times s}$ is a submatrix of $\widetilde{A}$ with columns of $\widetilde{A}_{S^*}$ equal to the columns of $\widetilde{A}$ on support set S^*, $(\widetilde{A}_{S^*})^{\#}$ standing for its pseudo-inverse and $I_{S^*} \in \mathbb{R}^{p \times s}$ is a submatrix of the identity matrix I_p with columns of I_{S^*} equal to the columns of I_p on support set S^*. Note that this estimator functions as an oracle, since in general the support set S^* is unknown. The oracle least squares estimate estimate satisfies the following result:

Proposition 4.4. *Assume that $S^* = \operatorname{supp}(x^*)$ is known and $s = |S^*|$. Under (4.8), $\widehat{x}^{\text{OLS}}$, the least square estimate of x^* on S^* satisfies*

$$\sum_{k \in S^*} (\widetilde{A}^\top (\widetilde{Y} - \widetilde{A}x^*))_k^2 \lesssim \|\widehat{x}^{\text{OLS}} - x^*\|_2^2 \lesssim \sum_{k \in S^*} (\widetilde{A}^\top (\widetilde{Y} - \widetilde{A}x^*))_k^2$$

with probability exceeding $1 - C/p$ for a universal positive constant C.

It's quite natural to compare our LASSO estimates with the OLS estimate. In the same spirit as [33], the OLS estimate can then be viewed as a benchmark and, in some sense, the previous result shows the optimality of the weighted LASSO estimator $\widehat{x}^{\text{WL}}$. Specifically, recall from Proposition 2.1 and (2.8), in the sparse setting $\|x^* - \widehat{x}^{\text{WL}}\|_2^2 \lesssim_\gamma \sum_{k \in S^*} d_k^2$ and $\|x^* - \widehat{x}^{\text{LASSO}}\|_2^2 \lesssim_\gamma sd^2$. By choosing each d_k to be as close as possible to $(\widetilde{A}^\top (\widetilde{Y} - \widetilde{A}x^*))_k$, we ensure that the risk of the weighted LASSO estimator $\widehat{x}^{\text{WL}}$ is as close as possible to that of the OLS estimate. In contrast, since d is a bound on $\max_k (\widetilde{A}^\top (\widetilde{Y} - \widetilde{A}x^*))_k$, the classical LASSO estimator $\widehat{x}^{\text{LASSO}}$ may exhibit higher errors than the OLS.

4.5 Comparison with previous rate results for Poisson CS

We note that the rates above are similar in spirit to the rates derived in a similar setting for estimators based on minimizing a regularized negative log-likelihood. Specifically, we compare our rates to those derived in [12]. In that work, the authors consider observations

$$Z \sim \mathcal{P}(TBD\theta^*),$$

where $T > 0$ is a scalar parameter controlling the expected total number of photons collected, B is a normalized, non-negative sensing matrix reflecting the physical constraints of optical systems, D is an orthonormal sparsifying basis, and θ^* is an s-sparse set of basis coefficients. [12] focused on the case where $\|x^*\|_1 = \|D\theta^*\|_1 = 1$ so that T alone reflected the overall signal strength, and where one of the basis vectors in D corresponded to a constant vector. The authors noted that the squared error of their penalized likelihood estimator scaled like $\frac{s \log p}{T}$ for T sufficiently large.

To compare that work with ours, we assume that the results in [12] generalize to the case where the sparsifying basis D is the identity matrix. We also focus on the case where $q = 1/2$ and the sensing matrix B is generated as follows:[5]

$$B_{i,j} = \frac{\widetilde{A}_{i,j}}{2\sqrt{n}} + \frac{1}{2n}, \qquad i = 1, \dots, n;\ j = 1, \dots, p.$$

From here we can see that if $T = n$, then $TB = A$. As a result, the rate of $\frac{s \log p}{T}$ in [12] is on the same order as the bound (4.13), $\frac{\|x^*\|_1 (\frac{1}{q} + s) \log p}{n}$, when $\|x^*\|_1 = 1$, $q = 1/2$, and $n \gtrsim \log p$.

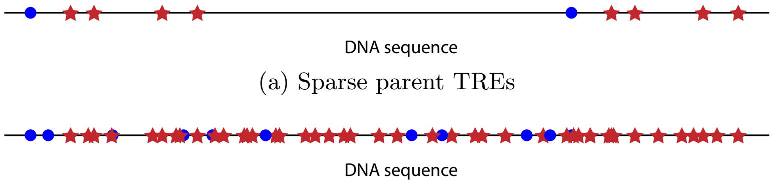
[5] In [12], $B_{i,j} = \frac{\widetilde{A}_{i,j}}{4\sqrt{n}} + \frac{3}{4n}$; that variation ensures Poisson intensities are bounded away from zero to facilitate analysis of the Poisson log-likelihood. We assume the work of [12] generalizes to the setting described in the text.

5 Case study: Poisson random convolution in genomics

This set-up is much less widely studied than the previous one, but it was the motivating problem at the origin of the present article. Indeed, this specific random convolution model is a toy model for bivariate Hawkes models or more precise Poissonian interaction functions [35, 7, 46]. Those point processes models have been used in neuroscience (spike train analysis) to model excitation from on neuron on another one or in genomics to model distance interaction along the DNA between motifs or occurrences of any kind of Transcription Regulatory Elements (TRE) [47, 48]. All the methods proposed in those articles assume that there is a finite "horizon" after which no interaction is possible (i.e. the support of the interaction function is finite and much smaller that the total length of the data) and so the corresponding inverse problem is well-posed. However, and in particular in genomics, it is not at all clear that such a horizon exists. Indeed, it is usually assumed that the interaction stops after 10000 bases. However, the 3D structure of DNA makes long-range "linear distances" on the DNA strand potentially irrelevant. An important question receiving increased attention is whether, if one had access to real 3D positions (and there is ongoing work to measure these positions), would it be possible to estimate the interaction functions without any assumption on its support?

The problem described here is a clear simplification of this complex problem, which in fact depends on the DNA fold: we restrict ourselves to the case where the DNA strand is just modeled by a circle (which is topologically reasonable for certain genomes) and the observations of TREs are binned. We wish to understand whether long range dependencies can be recovered once sparsity is assumed. Specifically, we formulate this problem as a sparse Poisson deconvolution problem.

Our model considers the interdependencies between two different kind of locations; for example, the first might be a transcription factor binding site (TFBS) and the second might be a transcription start site (TSS). Borrowing from the point process terminology, we call the first set of occurrences "parents" and the second set of occurrences "children". We assume the locations of the parents along the genome follow a uniform distribution, and each parent independently generates children in the surrounding genome according to the same distribution centered around the parent. This idea is illustrated in Figure 1.



(a) Sparse parent TREs

(b) Dense parent TREs

Figure 1: Illustration of the parents/children model. The locations of parents are shown as blue dots along the DNA sequence, and the locations of children generated by each parent are shown as red stars following the parent. The distribution of children is the same after each parent, but the parentage of each child becomes less clear as the density of parents increases.

More formally, let $U_1, \ldots, U_m$ be a collection of m i.i.d. realizations of a uniform random variable on the set $\{0, \ldots, p-1\}$; these corresponds to the parents' locations. Each parent U_i give birth independently to some children. The number of children of U_i at distance j is given by $N^i_{U_i+j}$, which is a Poisson variable with distribution $\mathcal{P}(x^*_j)$, where x^*_j represents the likelihood of a child's existence at distance j from its parent. Here we understand $U_i + j$ in a cyclic way, i.e. this is actually $U_i + j$ modulus p. We observe at each position k between 0 and $p-1$ the total number of children regardless of who their parent might be, *i.e.,*

$$Y_k = \sum_{i=1}^{m} N_k^i.$$

This problem can be translated as follows: we observe the U_i's and the Y_k's, whose conditional distribution given the U_i's is

$$Y_k \sim \mathcal{P}\left(\sum_{i=1}^{m} x^*_{k-U_i}\right).$$

Given the U_i's, the elements of $Y = (Y_0, \ldots, Y_{p-1})^\top$ are independent. The aim is to recover the vector $x^* = (x_0^*, \ldots, x_{p-1}^*)^\top$. We assume throughout that x^* is s-sparse for some $1 \leq s < \min(m, p)$. The sparsity is a reasonable assumption in genomics because linked parents and children in our model correspond to distinct chemical reactions in the underlying biochemical system.

The above model actually amounts to a random convolution (we are convolving the signal x^* by the random empirical measure of the parents). To the best of our knowledge, the analysis of such a convolution problem is entirely new. Other authors have studied random convolution, notably [49, 50, 51, 52], but those analyses do not extend to the problem considered here. For example, Candès and Plan [52, p5] consider random convolutions in which they observe a random subset of elements of the product Ax^*. They note that A is an isometry if the Fourier components of any row of A have coefficients with the same magnitude. In contrast, in our setting the Fourier coefficients of A are random and do not have uniform magnitude, and so the analysis in [52] cannot be directly applied in our setting. In particular, the ratio of p (the number of elements in x^* and the number of measurements) to m (the number of uniformly distributed parents) will play a crucial role in our analysis but is not explored in the existing literature.

5.1 Poisson random convolution model

Let us introduce the multinomial variable $\mathbb{N}$, defined for all $k \in \mathbb{Z}$ by

$$\mathbb{N}(k) = \text{card}\left\{i : \ U_i = k[p]\right\}, \tag{5.1}$$

where $k[p]$ denotes k modulo p. It represents the number of parents at position k on the circle. Note that $\sum_{u=0}^{p-1} \mathbb{N}(u) = m$, a fact which will be extensively used in proofs. Let us denote

$$A = \begin{pmatrix} \mathbb{N}(0) & \mathbb{N}(p-1) & \cdots & \mathbb{N}(1) \\ \mathbb{N}(1) & \mathbb{N}(0) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \mathbb{N}(p-1) \\ \mathbb{N}(p-1) & \cdots & \mathbb{N}(1) & \mathbb{N}(0) \end{pmatrix},$$

where $A_{\ell,k} = \mathbb{N}(\ell - k)$. Using this notation, we have in fact the observation model

$$Y \sim \mathcal{P}(Ax^*).$$

Note that $\mathbb{E}(A) = m\mathbb{1}_p\mathbb{1}_p^\top$. Therefore, in expectation, all its eigenvalues are null except the first one. In this sense it is a badly ill-posed problem despite the fact that the sensing matrix is square here. The ill-posedness can also be viewed by the fact that in expectation, we are convolving the unknown x^* with a uniform distribution, which is known to be an unsolvable problem. Therefore (as in many works on compressed sensing), we rely on the randomness to prove that Assumption RE(κ_1, κ_2) is satisfied with high probability.

Finally, since m corresponds to the number of parents, it controls the total number of (randomly shifted) copies of x^* that are observed and hence determines the expected total number of observed events, which is $m\|x^*\|_1$. In this sense, m controls the Poisson rates and signal-to-noise ratio, and also controls the ill-posedness of the inverse problem.

5.2 Rescaling and recentering

As for the Bernoulli case, we first rescale and recenter the sensing matrix

$$\widetilde{A} := \frac{1}{\sqrt{m}}A - \frac{\sqrt{m}-1}{p}\mathbb{1}\mathbb{1}^\top,$$

which satisfies that $\mathbb{E}(\widetilde{A}^\top \widetilde{A}) = I_p$. Moreover for any $k \in \{0, \ldots, p-1\}$, we can easily define:

$$\widetilde{Y}_k := \frac{1}{\sqrt{m}}Y_k - \frac{\sqrt{m}-1}{p}\overline{Y},$$

where $\overline{Y} = \frac{1}{m}\|Y\|_1$. Note that because of the particular form of A, $\mathbb{E}[\widetilde{Y}|\widetilde{A}] = \widetilde{A}x^*$, (see Lemma D.1 in the Appendix) which explains why in this case the remainder term r_k described in Section 3 is actually null.

5.3 Assumption RE holds with high probability

Let $\widetilde{G} := \widetilde{A}^\top \widetilde{A}$.

Proposition 5.1. *There exists absolute positive constants κ and C and an event of probability larger than $1 - C/p$, on which*

$$\forall k, \ell, \left| \left(\widetilde{G} - I_p\right)_{k,\ell} \right| \leq \xi := \kappa \left(\frac{\log p}{\sqrt{p}} + \frac{\log^2 p}{m} \right). \tag{5.2}$$

This implies on the same event that Assumption RE(κ_1, κ_2) holds with

$$\kappa_1 = \sqrt{\xi} \quad \text{and} \quad \kappa_2 = 1.$$

This result is a result of concentration inequalities for U-statistics (see Proposition D.1 in the appendix).

5.4 Choice of the weights

As in the Bernoulli case (Section 4), we consider both a constant weight (corresponding to a classical LASSO estimator) and non-constant weights that follow from Lemma 3.1.

5.4.1 Definition of constant weight and rates for the estimate $\widehat{x}^{\mathrm{LASSO}}$

For all $k = 1, \ldots, p$, define the vector $V_k \in \mathbb{R}^p$ so that the ℓ^{th} element is

$$V_{k,\ell} := \left(\frac{\mathbb{N}(\ell - k) - \frac{m-1}{p}}{m} \right)^2 . \tag{5.3}$$

Let

$$W = \max_{k \in \{1,\ldots,p\}} \langle V_k, \mathbb{N} \rangle .$$

Furthermore, let

$$B = \max_{u \in \{0,\ldots,p-1\}} \frac{1}{m} \left| \mathbb{N}(u) - \frac{m-1}{p} \right| = \|V_1\|_\infty^{1/2} .$$

Then choose the following constant weight

$$d := \sqrt{4W \log p} \left[\sqrt{\overline{Y} + \frac{5 \log p}{3m}} + \sqrt{\frac{\log p}{m}} \right] + \frac{2B \log p}{3} , \tag{5.4}$$

and one can prove that it satisfies Assumption Weights(d) except on an event of probability of order $1/p$ (see Proposition D.2).

The order of magnitude of d is given by Proposition D.3 in the appendix and states that

$$d^2 \lesssim \left(\frac{\log(p)^2}{p} + \frac{\log(p)^3}{m} \right) \left(\|x^*\|_1 + \frac{\log(p)}{m} \right) .$$

We consider once again a sparse signal x^* with support size s. By fixing $\varepsilon = 1/2$ and $\gamma > 2$, it is easy to see that (2.6) is implied for large p by

$$s = o \left(\min \left(\frac{\sqrt{p}}{\log(p)}, \frac{m}{\log(p)^2} \right) \right) , \tag{5.5}$$

and applying Proposition 2.1 gives the following proposition:

Proposition 5.2. *For the weight defined in Equation* (5.4)*, if* (5.5) *holds, then except on an event of probability of order* $1/p$*,*

$$\|\widehat{x}^{\mathrm{LASSO}} - x^*\|_2^2 \lesssim_\gamma s \left(\frac{\log(p)^2}{p} + \frac{\log(p)^3}{m} \right) \left(\|x^*\|_1 + \frac{\log(p)}{m} \right) . \tag{5.6}$$

5.4.2 Definition of non-constant weights and rates for the estimate $\widehat{x}^{\mathrm{WL}}$

For all $k = 0, \ldots, p-1$, one can choose the non-constant weights given by

$$d_k = \sqrt{4 \log p} \left[\sqrt{\langle V_k, Y \rangle + \frac{5B^2 \log p}{3}} + \sqrt{B^2 \log p} \right] + \frac{2B \log p}{3} , \tag{5.7}$$

with for all k in $\{0, \ldots, p-1\}$. These weights satisfy Assumption Weights(d) except on an event of probability of order $1/p$ (see Proposition D.4). The order of magnitude of the d_k's are subtle to derive and we have been able to derive them only in the range

$$\sqrt{p} \log p \lesssim m \lesssim \frac{p}{\log(p)} , \tag{5.8}$$

where Proposition D.5 shows that

$$\frac{x_k^* \log p}{m} + \frac{\log p}{p} \sum_{u \neq k} x_u^* + \frac{\log^2 p}{m^2} \lesssim d_k^2 \lesssim \frac{x_k^* \log p}{m} + \frac{\log^2 p}{p} \sum_{u \neq k} x_u^* + \frac{\log^4 p}{m^2}.$$

We consider a sparse signal x^* with support S^*. Then, since $m \lesssim p$,

$$\frac{\|d_{S^*}\|_2^2}{d_{\min}^2} \lesssim \frac{\|x^*\|_1 \left(\frac{\log p}{m} + \frac{s \log^2 p}{p}\right) + \frac{s \log^4 p}{m^2}}{\frac{\log p}{p}\|x^*\|_1 + \frac{\log^2 p}{m^2}} \lesssim \frac{p}{m} + s \log^2 p.$$

Since $\xi \simeq \log(p)/\sqrt{p}$ in the range (5.8), one can then easily see that for fixed $\gamma > 2$ and $\varepsilon = 1/2$, (2.4) is implied for large p by

$$s = o\left(\frac{\sqrt{p}}{\log^3 p}\right). \tag{5.9}$$

This condition is equivalent up to logarithmic factors to (5.5) which is necessary for the classical LASSO as soon as (5.8) holds.

It remains to apply Proposition 2.1 to obtain the following proposition.

Proposition 5.3. *For the weights defined in Equation (5.7), if (5.8) and (5.9) hold, then except on an event of probability of order $1/p$,*

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_2^2 \lesssim \left(\frac{\|x^*\|_1 \log p}{m} + \frac{s\|x^*\|_1 \log^2 p}{p} + \frac{s \log^4 p}{m^2}\right). \tag{5.10}$$

If the signal is strong enough (i.e., $\|x^*\|_1 = \omega\left(\log(p)/m\right)$) and in the range (5.8), the bound on the risk of classical LASSO estimator, (5.6), is of order $\frac{s\|x^*\|_1 \log^3 p}{m}$. The ratio of the two bounds,

$$\frac{\frac{\|x^*\|_1 \log p}{m} + \frac{s\|x^*\|_1 \log^2 p}{p} + \frac{s \log^4 p}{m^2}}{\frac{s\|x^*\|_1 \log^3 p}{m}} = \frac{1}{s \log^2 p} + \frac{m}{p \log p} + \frac{\log p}{m\|x^*\|_1}$$

illustrates that the upper bound on the weighted LASSO estimator is much smaller as p tends to infinity, if $\|x^*\|_1 = \omega\left(\log(p)/m\right)$, (5.8) and (5.9) hold.

5.5 Comparison with the oracle least squares estimate

We consider the *oracle least squares estimate* on S^*, the true signal support:

$$\widehat{x}^{\mathrm{OLS}} := I_{S^*}(\widetilde{A}_{S^*})^{\#}\widetilde{Y}. \tag{5.11}$$

See Section 4.4 for definitions of the notation in this estimator. As before, this estimator functions as an oracle, since in general the support set S^* is unknown. The oracle least squares estimate estimate satisfies the following result:

Proposition 5.4. *Assume that $S^* = \mathrm{supp}(x^*)$ is known and $s = |S^*|$. Under (5.5), then $\widehat{x}^{\mathrm{OLS}}$, the least square estimate of x^* on S^* satisfies*

$$\sum_{k \in S^*} (\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k^2 \lesssim \|\widehat{x}^{\mathrm{OLS}} - x^*\|_2^2 \lesssim \sum_{k \in S^*} (\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k^2$$

with probability exceeding $1 - C/p$ for universal positive constants C.

See the discussion on the implications of this in Section 4.4, which apply in this setting.

5.6 Simulations

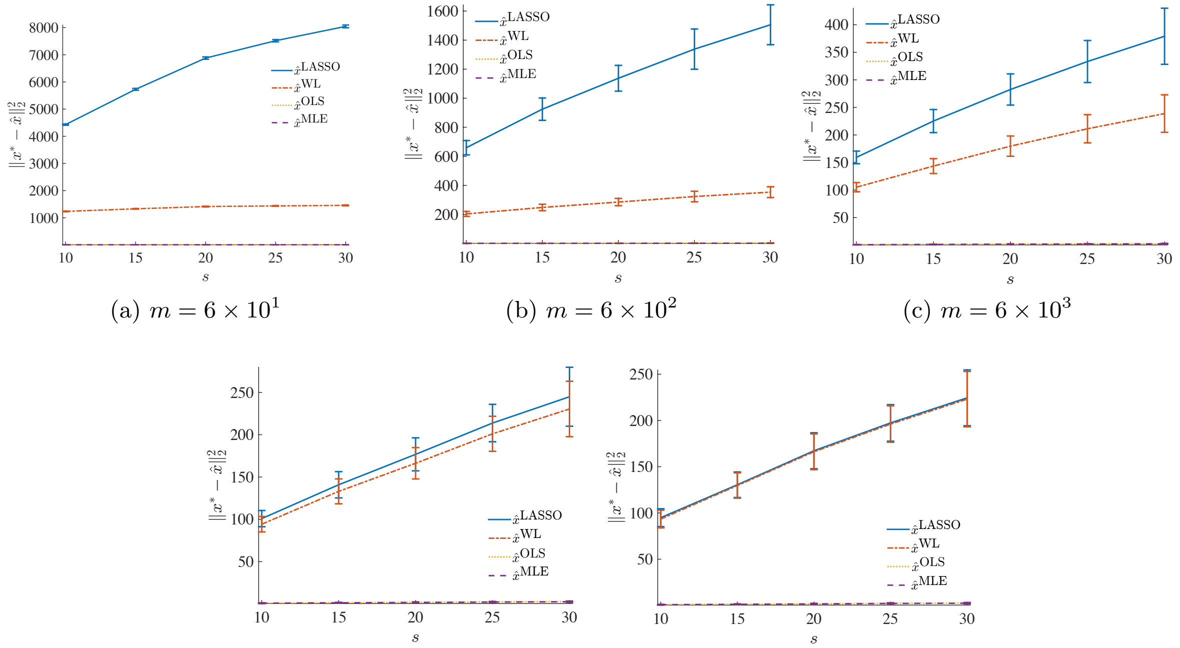
In this section we simulate the random convolution model described above and the performance of the (unweighted) LASSO and weighted LASSO estimators. We compare the performance of these estimators to an oracle maximum likelihood estimator (MLE) estimated over the true support S^* which is denoted by $\widehat{x}^{\text{MLE.}}$ Note that this MLE is efficient and therefore should be the estimate with minimum variance among unbiased estimators and have the minimum risk in some sense.

We have shown that when m is small relative to p, for both weighted LASSO and least-squares estimators, the MSE upper bound scales like $\frac{\|x^*\|_1 \log p}{m}$; for the LASSO estimator, the MSE scales like $\frac{s\|x^*\|_1 \log^3 p}{m}$; in the below, we present a simulation showing that these upper bounds are tight. First we examine the MSE of the LASSO and weighted LASSO estimators as a function of s, the sparsity level of x^* for various m. We set $p = 5000$ and s ranges from 10 to 30. The A matrix is randomly generated at each experiment. The x^* is fixed for each value of s, and $\|x^*\|_1$ is kept the same for different s values. The tuning parameter $\gamma = 2.01$ such that it satisfies the constraint $\gamma > 2$. Each point in the plots is averaged over 100 random realizations.

In Figure 2, we explore the MSE as a function of s for different values of m, with $p = 5000$ in the Poisson random convolution setting for genomics. For m in the range in (5.8), for fixed p, m, and $\|x^*\|_1$, our theory predicts that weighted LASSO outperform standard LASSO by a factor of s, as reflected in (a)-(d). Note that some of the values of m in these plots are outside the range considered in (5.8). Figures 2(a)-(b) satisfy the conditions of our theory and behave as expected; Figures 2(c)-(d) do not satisfy our conditions and hence do not demonstrate the gains predicted by our theory. Specifically, as m gets large, we see a greater dependence of the MSE on s for the weighted LASSO. This effect is predicted by the theory. The weights used by the Weighted LASSO in (5.7) depend on the variances in $\langle V_k, Y\rangle$, where V_k is defined in (5.3). As m grows, the elements of V_k become more uniform as a consequence of the strong law of large numbers, so the inner products (and hence the d_k's) all start to be close to the same value. When this happens, the weighted LASSO estimate closely approximates the classical LASSO estimate, as illustrated in the figure.

Next, we examine the MSE of the classical and weighted LASSO estimators as a function of p, the length of x^*. In this experiment, we set $s = 10$ and p ranges from 1000 to 10000. For each p, we set $m \propto \sqrt{p}\log(p)$ (m varies from 11 to 40 in this experiment). This specific choice of m is made due to the requirement of $m \gtrsim \sqrt{p}\log(p)$ in our rate results (5.8). The x^* is fixed for each of the p value, and $\|x^*\|_1$ is kept the same for different p. A is randomly generated in each trial. The tuning parameter γ is set to 2.01 such that it satisfies the constraint $\gamma > 2$. Each point in the plots is averaged over 100 random realizations.

Figure 3a shows MSE as a functions of p. The weighted LASSO estimators outperforms the classical LASSO estimators. Note that because of our choice of m, our theorem predicts that the MSE of weighted LASSO estimator scales like $\frac{\|x^*\|_1}{\sqrt{p}}$; the MSE of LASSO scales like $\frac{s\|x^*\|_1 \log^2(p)}{\sqrt{p}}$. With fixed s and $\|x^*\|_1$, the weighted LASSO estimator has an error rate $\propto 1/\sqrt{p}$, while the error rate of the LASSO estimator $\propto \log(p)^2/\sqrt{p}$. To better show the relationship between the MSE and $1/\sqrt{p}$, we plot two additional lines $\propto 1/\sqrt{p}$ for the one-step estimators $\widehat{x}^{\text{WL}}$ and $\widehat{x}^{\text{LASSO}}$ in Figure 3a. In Figure 3a, the MSE curve of $\widehat{x}^{\text{WL}}$ follows the $6.5 \times 10^5/\sqrt{p}$ curve almost perfectly, while the MSE curve of $\widehat{x}^{\text{LASSO}}$ decreases (with p) more slowly than the $1.3 \times 10^6/\sqrt{p}$ curve. This shows that the MSE of $\widehat{x}^{\text{WL}}$ has a rate $\propto 1/\sqrt{p}$,



(a) $m = 6 \times 10^1$

(b) $m = 6 \times 10^2$

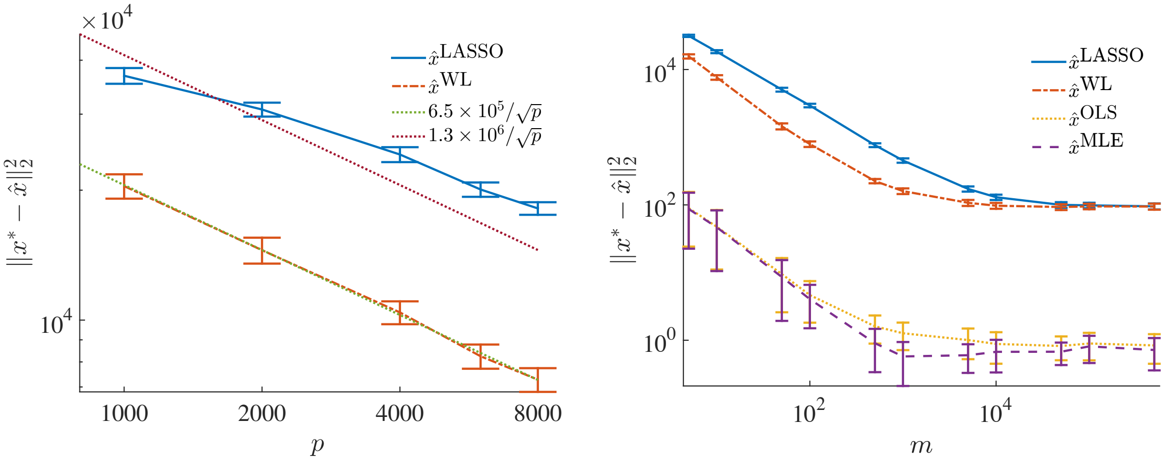
(c) $m = 6 \times 10^3$

(d) $m = 6 \times 10^4$

(e) $m = 6 \times 10^5$

Figure 2: MSE vs s for different values of m, with $p = 5000$ in the Poisson random convolution setting for genomics. This simulation compares the estimators oracle least squares (5.11), classical (unweighted) LASSO (1.2) with d in (5.4), and the weighted LASSO (1.3) with weights in (5.7). The oracle maximum likelihood estimators yield the smallest estimation error, and the oracle least squares estimator yields smaller estimation error than the LASSO and weighted LASSO estimators. The weighted LASSO estimator outperforms the standard LASSO estimator, as predicted by the theory. Note that some of the values of m in these plots are outside the range considered in (5.8). For m satisfying (5.8), the error of the weighted LASSO does not scale with s, as predicted by the theory. As m gets large and exceeds this range, however, we see a greater dependence of the MSE on s for the weighted LASSO. This effect is predicted by the theory. The weights used by the weighted LASSO in (5.7) depend on the variances in $\langle V_k, Y \rangle$, which become more uniform as m gets large.

while the MSE of $\widehat{x}^{\mathrm{LASSO}}$ is slower than $1/\sqrt{p}$, as predicted by the theory. Figure 3b shows MSE vs m for $p = 5000$, where results are averaged over 50 trials. This plot demonstrates that for large m, the weighted LASSO and classical LASSO are nearly equivalent, while for the range of m in (5.8), the Weighted LASSO has lower errors. In view of Figures 2 and 3b, it may be possible to drop the restriction on the left hand side of Inequality (5.8). However, it is unlikely that the restriction on the right hand side can be dropped.



(a) MSE vs. p

(b) MSE vs. m

Figure 3: Simulation results. 3a shows MSE vs. p with $m \propto \sqrt{p}\log p$ for the weighted LASSO and standard LASSO estimators. Weighted LASSO outperforms standard LASSO. The oracle estimators $\widehat{x}^{\mathrm{OLS}}$ and $\widehat{x}^{\mathrm{MLE}}$ perform similarly. The MSE curve of $\widehat{x}^{\mathrm{WL}}$ follows the $6.5\times10^5/\sqrt{p}$ curve almost perfectly, while the MSE curve of $\widehat{x}^{\mathrm{LASSO}}$ decreases (with p) more slowly than the $1.3\times10^6/\sqrt{p}$ curve, showing that $\widehat{x}^{\mathrm{WL}}$ has a rate $\propto 1/\sqrt{p}$, while $\widehat{x}^{\mathrm{LASSO}}$ has a rate slower than $1/\sqrt{p}$. 3b shows MSE vs m for $p = 5000$ in the Poisson random convolution setting for genomics. Results averaged over 50 trials. This plot demonstrates that for large m, the weighted LASSO and classical LASSO are nearly equivalent, while for the range of m in (5.8), the Weighted LASSO has lower errors.

6 Discussion and Conclusions

The data-dependent weighted LASSO method presented in this paper, and based on the general recipe stated in Section 3.1, is a novel approach to sparse inference in the presence of heteroscedastic noise. We show that using concentration inequalities to learn data-dependent weights leads to estimation errors which closely approximate errors achievable by an oracle with knowledge of the true signal support. To use this technique, concentration inequalities which account for the noise distribution are used to set data-dependent weights which satisfy the necessary assumptions with high probability.

In contrast to earlier work on sparse Poisson inverse problems [12], the estimator proposed here is computationally tractable. In addition, earlier analyses required ensuring that the product Ax^* was bounded away from zero, which limited the applicability of the analysis. Specifically, the random convolution problem described in Section 5 could not be directly analyzed using the techniques described in [12].

Our technique can also yield immediate insight into the role of background contamination. Consider a setting in which we observe

$$Y \sim \mathcal{P}(Ax^* + b),$$

where $b \in \mathbb{R}_+^n$ is a known (typically constant) background intensity. In imaging, for instance, this would correspond to ambient light or dark current effects. While b contributes to the noise variance, it does not provide any information about the unknown signal x^*. Since b is known, it can easily be subtracted from the observations in the formation of $\widetilde{Y}$ and we can use exactly the estimation framework described above (*e.g.*, the estimator in (1.3)). However, because b impacts the variance of the observations, it will increase the value of v in Lemma 3.1, leading to a proportional increase in the weights and hence the ℓ_2 error decay rates. From here we can see that the error decay rates will increase linearly with the amount of background contamination.

7 Acknowledgments

We gratefully acknowledge the support of AFOSR award 14-AFOSR-1103, NSF award CCF-1418976, the University of Nice Sophia Antipolis Visiting Professor program and ANR 2011 BS01 010 01 projet Calibration. The research of Vincent Rivoirard benefited from the support of the *Chaire Economie et Gestion des Nouvelles Données*, under the auspices of Institut Louis Bachelier, Havas-Media and Paris-Dauphine. The authors thank the anonymous referee for insightful comments and suggestions.

References

[1] R. Willett and R. Nowak. Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging. *IEEE Trans. Med. Imag.*, 22(3):332–350, 2003.

[2] D. Lingenfelter, J. Fessler, and Z. He. Sparsity regularization for image reconstruction with Poisson data. In *IS&T/SPIE Electronic Imaging*, page 72460F. International Society for Optics and Photonics, 2009.

[3] T. Schmidt. Optimal image-based weighting for energy-resolved ct. *Medical physics*, 36(7):3018–3027, 2009.

[4] K. Borkowski, S. Reynolds, D. Green, U. Hwang, R. Petre, K. Krishnamurthy, and R. Willett. Nonuniform expansion of the youngest galactic supernova remnant g1.9+0.3. *The Astrophysical Journal Letters*, 790(2), 2014. arXiv:1406.2287.

[5] K. Borkowski, S. Reynolds, D. Green, U. Hwang, R. Petre, K. Krishnamurthy, and R. Willett. Supernova ejecta in the youngest galactic supernova remnant g1.9+0.3. *The Astrophysical Journal Letters*, 771(1), 2013. arXiv:1305.7399.

[6] J.-L. Starck and J. Bobin. Astronomical data analysis and sparsity: from wavelets to compressed sensing. *Proc. IEEE*, 98(6):1021–1030, 2010.

[7] L. Sansonnet. Wavelet thresholding estimation in a poissonian interactions model with application to genomic data. *Scandinavian Journal of Statistics*, 41(1):200–226, 2014.

[8] C. Estan and G. Varghese. New directions in traffic measurement and accounting: Focusing on the elephants, ignoring the mice. *ACM Transactions on Computer Systems*, 21(3):270–313, 2003.

[9] Y. Lu, A. Montanari, B. Prabhakar, S. Dharmapurikar, and A. Kabbani. Counter Braids: A novel counter architecture for per-flow measurement. *ACM SIGMETRICS Performance Evaluation Review*, 36(1):121–132, 2008.

[10] D. Osgood. Poisson-based regression analysis of aggregate crime rates. *Journal of Quantitative Criminology*, 16(1), 2000.

[11] J.-M. Xu, A. Bhargava, R. Nowak, and X. Zhu. Socioscope: Spatio-temporal signal recovery from social media. *Machine Learning and Knowledge Discovery in Databases*, 7524, 2012.

[12] X. Jiang, G. Raskutti, and R. Willett. Minimax optimal rates for poisson inverse problems with physical constraints. *IEEE Trans. Inf. Theory*, 61(8):4458–4474, 2015. arXiv:1403:6532.

[13] M. Raginsky, R. Willett, Z. Harmany, and R. Marcia. Compressed sensing performance bounds under poisson noise. *IEEE Trans. Signal Process.*, 58(8):3990–4002, 2010.

[14] I. Rish and G. Grabarnik. Sparse signal recovery with exponential-family noise. In *Compressed Sensing & Sparse Filtering*, pages 77–93. Springer, 2014.

[15] M. Raginsky, S. Jafarpour, Z. Harmany, R. Marcia, R. Willett, and R. Calderbank. Performance bounds for expander-based compressed sensing in Poisson noise. *IEEE Trans. Signal Process.*, 59(9), 2011. arXiv:1007.2377.

[16] M. Rohban, V. Saligrama, and D. Vaziri. Minimax optimal sparse signal recovery with Poisson statistics. *IEEE Trans. Signal Process.*, 64(13):3495–3508, 2016.

[17] S. Ivanoff, F. Picard, and V. Rivoirard. Adaptive lasso and group-lasso for functional poisson regression. *Journal of Machine Learning Research*, 17(55):1–46, 2016.

[18] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. Roy. Statist. Soc. Ser. B*, 58(1):267–288, 1996.

[19] E. J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory*, 52(12):5406–5425, 2006.

[20] D. L. Donoho. Compressed sensing. *IEEE Trans. Inf. Theory*, 52(4):1289–1306, 2006.

[21] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and dantzig selector. *Ann. Statist.*, pages 1705–1732, 2009.

[22] F. Bunea, A. Tsybakov, and M. Wegkamp. Aggregation for Gaussian regression. *Ann. Statist.*, 35(4):1674–1697, 2007.

[23] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, 2:90–102, 2008.

[24] S. Van de Geer. High-dimensional generalized linear models and the LASSO. *Ann. Statist.*, 36(2):614–645, 2008.

[25] P. Bühlmann and S. van de Geer. *Statistics for high-dimensional data*. Springer Series in Statistics. Springer, Heidelberg, 2011. Methods, theory and applications.

[26] D. Paul, E. Bair, T. Hastie, and R. Tibshirani. "Preconditioning" for feature selection and regression in high-dimensional problems. *Ann. Statist.*, 36(4):1595–1618, 2008.

[27] F. Wauthier, N. Jojic, and M. Jordan. A comparative framework for preconditioned Lasso algorithms. In *Advances in Neural Information Processing Systems*, pages 1061–1069, 2013.

[28] J. Huang and N. Jojic. Variable selection through correlation sifting. In Vineet Bafna and S. Cenk Sahinalp, editors, *Research in Computational Molecular Biology*, pages 106–123, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[29] S. van de Geer, P. Bühlmann, and S. Zhou. The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Electronic Journal of Statistics*, 5:688–749, 2011.

[30] A. Juditsky and A. Nemirovski. Accuracy guarantees for ℓ_1-recovery. *IEEE Trans. Inform. Theory*, 57(12):7818–7839, 2011.

[31] E. J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, May 2008.

[32] S. Negahban, B. Yu, M. Wainwright, and P. Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

[33] E. J. Candès and T. Tao. The dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, pages 2313–2351, 2007.

[34] K. Bertin, E. Le Pennec, and V. Rivoirard. Adaptive dantzig density estimation. *Annales de l'institut Henri Poincaré (B)*, 47:43–74, 2011.

[35] N. Hansen, P. Reynaud-Bouret, and V. Rivoirard. Lasso and probabilistic inequalities for multivariate point processes. *Bernoulli*, 21(1):83–143, 2015.

[36] J. Huang, S. Ma, and C.-H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, 18(4):1603, 2008.

[37] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[38] D. L. Donoho and I. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455, 1994.

[39] A. Juditsky and S. Lambert-Lacroix. On minimax density estimation on $\mathbb{R}$. *Bernoulli*, 10(2):187–220, 2004.

[40] P. Reynaud-Bouret and V. Rivoirard. Near optimal thresholding estimation of a poisson intensity on the real line. *Electronic journal of statistics*, 4:172–238, 2010.

[41] P. Reynaud-Bouret, V. Rivoirard, and C. Tuleau-Malot. Adaptive density estimation: a curse of support? *Journal of Statistical Planning and Inference*, 141(1):115–139, 2011.

[42] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single pixel imaging via compressive sampling. *IEEE Sig. Proc. Mag.*, 25(2):83–91, 2008.

[43] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical learning with sparsity: the Lasso and generalizations.* CRC press, 2015.

[44] G. Raskutti, M. Wainwright, and B. Yu. Restricted eigenvalue properties for correlated gaussian designs. *Journal of Machine Learning Research*, 11(Aug):2241–2259, 2010.

[45] Y. Li and G. Raskutti. Minimax optimal convex methods for Poisson inverse problems under ℓ_q-ball sparsity. *arXiv:1604.08943*, 2016.

[46] L. Sansonnet and C. Tuleau-Malot. A model of Poissonian interactions and detection of dependence. *Stat. Comput.*, 25(2):449–470, 2015.

[47] G. Gusto and S. Schbath. FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Stat. Appl. Genet. Mol. Biol.*, 4:Art. 24, 28 pp. (electronic), 2005.

[48] L. Carstensen, A. Sandelin, O. Winther, and N. Hansen. Multivariate hawkes process models of the occurrence of regulatory elements. *BMC bioinformatics*, 11(1):456, 2010.

[49] J. Romberg. Compressive sensing by random convolution. *SIAM Journal on Imaging Sciences*, 2(4):1098–1128, 2009.

[50] R. Marcia and R. Willett. Compressive coded aperture superresolution image reconstruction. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 833–836. IEEE, March 2008.

[51] Z. Harmany, R. Marcia, and R. Willett. Spatio-temporal compressed sensing with coded apertures and keyed exposures. *arXiv:1111.7247*, 2011.

[52] E. J. Candès and Y. Plan. A probabilistic and RIPless theory of compressed sensing. *IEEE Trans. Inf. Theory*, 57(11):7235–7254, 2011.

[53] P. Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

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

[55] C. Houdré and P. Reynaud-Bouret. Exponential inequalities, with constants, for u-statistics of order two. In *Stochastic inequalities and applications*, pages 55–69. Springer, 2003.

[56] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*, volume 40. Cambridge University Press, 2015.

[57] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. Uniform uncertainty principle for bernoulli and subgaussian ensembles. *Constructive Approximation*, 28(3):277–289, 2008.

[58] E. Giné, R. Latała, and J. Zinn. Exponential and moment inequalities for U-statistics. In *High dimensional probability, II (Seattle, WA, 1999)*, volume 47 of *Progr. Probab.*, pages 13–38. Birkhäuser Boston, Boston, MA, 2000.

A Proofs of the LASSO bounds of Section 2

The weighted LASSO regularizer, $\|Dx\|_1$ is *decomposable* in the sense of [32], and hence we can leverage variants of the results developed in that paper to characterize the performance of the weighted LASSO estimator. The results of [32] applied naïvely to the proposed estimator results in rates equivalent to those associated with the classical (unweighed) LASSO. Thus several technical details of the analysis must be adjusted to derive tight bounds for the weighted LASSO estimator. This section contains those details for the sake of completeness.

Before proving Proposition 2.1, we need the following supporting lemma.

Lemma A.1. *Consider any optimal solution $\widehat{x}^{\mathrm{WL}}$ to the weighted LASSO optimization problem with $\gamma > 2$. Then, under Assumption Weights($\{d_k\}_k$), for any $S \subseteq \{1, \ldots, p\}$, the error $\Delta = \widehat{x}^{\mathrm{WL}} - x^*$ satisfies*

$$\|D\Delta_{S^c}\|_1 \leq \frac{\gamma+2}{\gamma-2}\|D\Delta_S\|_1 + \frac{2\gamma}{\gamma-2}\|Dx^*_{S^c}\|_1.$$

Remark A.1. *Lemma A.1 implies*

$$\|D\Delta\|_1 = \|D\Delta_S\|_1 + \|D\Delta_{S^c}\|_1 \leq \frac{2\gamma}{\gamma-2}\|D\Delta_S\|_1 + \frac{2\gamma}{\gamma-2}\|Dx^*_{S^c}\|_1. \tag{A.1}$$

Proof of Lemma A.1. Let $\Delta := \widehat{x}^{\mathrm{WL}} - x^*$. Our proof follows the structure of Lemma 3 of [32]:

$$\begin{aligned}
\|\widetilde{Y} - \widetilde{A}\widehat{x}^{\mathrm{WL}}\|_2^2 - \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 =& \|\widetilde{Y} - \widetilde{A}x^* - \widetilde{A}\Delta\|_2^2 - \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 \\
=& \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 - 2\langle \widetilde{Y} - \widetilde{A}x^*, \widetilde{A}\Delta\rangle + \|\widetilde{A}\Delta\|_2^2 - \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 \\
=& -2\langle \widetilde{Y} - \widetilde{A}x^*, \widetilde{A}\Delta\rangle + \|\widetilde{A}\Delta\|_2^2 \\
\geq& -2|\langle D^{-1}\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*), D\Delta\rangle| \\
\geq& -2\|D\Delta\|_1,
\end{aligned}$$

by Assumption Weights($\{d_k\}_k$). Now by the "basic inequality" we have

$$\|\widetilde{Y} - \widetilde{A}\widehat{x}^{\mathrm{WL}}\|_2^2 - \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 \leq \gamma\left[\|Dx^*\|_1 - \|D(x^* + \Delta)\|_1\right].$$

Therefore,

$$-2\|D\Delta\|_1 \leq \gamma\left[\|Dx^*\|_1 - \|D(x^* + \Delta)\|_1\right].$$

Since

$$\begin{aligned}\|D(x^*+\Delta)\|_1 &\geq \|D(x^*_S+\Delta_{S^c})\|_1 - \|D(x^*_{S^c}+\Delta_S)\|_1\\ &\geq \|Dx^*_S\|_1 + \|D\Delta_{S^c}\|_1 - \|Dx^*_{S^c}\|_1 - \|D\Delta_S\|_1,\end{aligned}$$

we obtain

$$-2\|D\Delta_S\|_1 - 2\|D\Delta_{S^c}\|_1 \leq \gamma\|Dx^*\|_1 - \gamma\left(\|Dx^*_S\|_1 + \|D\Delta_{S^c}\|_1 - \|Dx^*_{S^c}\|_1 - \|D\Delta_S\|_1\right)$$

and

$$\|D\Delta_{S^c}\|_1 \leq \frac{\gamma+2}{\gamma-2}\|D\Delta_S\|_1 + \frac{2\gamma}{\gamma-2}\|Dx^*_{S^c}\|_1.$$

□

Proof of Proposition 2.1. We still denote $\Delta := \widehat{x}^{\mathrm{WL}} - x^*$. By the "basic inequality" we have

$$\|\widetilde{Y} - \widetilde{A}\widehat{x}^{\mathrm{WL}}\|_2^2 - \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 \leq \gamma\left[\|Dx^*\|_1 - \|D(x^*+\Delta)\|_1\right].$$

This gives

$$-2\langle \widetilde{Y} - \widetilde{A}x^*, \widetilde{A}\Delta\rangle + \|\widetilde{A}\Delta\|_2^2 \leq \gamma\left[\|Dx^*\|_1 - \|D(x^*+\Delta)\|_1\right],$$

or

$$\begin{aligned}\|\widetilde{A}\Delta\|_2^2 \leq& 2\langle \widetilde{Y} - \widetilde{A}x^*, \widetilde{A}\Delta\rangle + \gamma\left[\|Dx^*\|_1 - \|D(x^*+\Delta)\|_1\right]\\ \leq& 2\|D^{-1}(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))\|_\infty\|D\Delta\|_1 + \gamma\left[\|Dx^*\|_1 - \|D(x^*+\Delta)\|_1\right]\\ \leq& 2\|D\Delta\|_1 + \gamma\left[\|Dx^*\|_1 - \|D(x^*+\Delta)\|_1\right]\\ =& 2\|D\Delta\|_1 + \gamma\|Dx^*_S\|_1 + \gamma\|Dx^*_{S^c}\|_1 - \gamma\|D(x^*+\Delta)_S\|_1 - \gamma\|D(x^*+\Delta)_{S^c}\|_1\\ \leq& 2\|D\Delta\|_1 + \gamma\|D\Delta_S\|_1 + \gamma\|Dx^*_{S^c}\|_1\\ \leq& \rho_\gamma\left(\|D\Delta_S\|_1 + \|Dx^*_{S^c}\|_1\right),\end{aligned}$$

where for the last step we use (A.1) and the definition of ρ_γ. We now provide a lower bound to the left hand side. First we use (A.1) to bound

$$\|\Delta\|_1 \leq \frac{\tilde{\rho}_\gamma}{d_{\min}}\left(\|D\Delta_S\|_1 + \|Dx^*_{S^c}\|_1\right) \leq \frac{\tilde{\rho}_\gamma}{d_{\min}}\left(\|d_S\|_2\|\Delta\|_2 + \|Dx^*_{S^c}\|_1\right),$$

where

$$\tilde{\rho}_\gamma := \frac{2\gamma}{(\gamma-2)} \quad \text{and} \quad \|d_S\|_2 := \left(\sum_{k\in S} d_k^2\right)^{1/2}.$$

Combining this with Assumption RE(κ_1,κ_2) gives

$$\begin{aligned}\|\widetilde{A}\Delta\|_2 \geq& \kappa_2\|\Delta\|_2 - \frac{\tilde{\rho}_\gamma\kappa_1}{d_{\min}}\left[\|d_S\|_2\|\Delta\|_2 + \|Dx^*_{S^c}\|_1\right]\\ =& \|\Delta\|_2\left(\kappa_2 - \frac{\tilde{\rho}_\gamma\kappa_1}{d_{\min}}\|d_S\|_2\right) - \frac{\tilde{\rho}_\gamma\kappa_1}{d_{\min}}\|Dx^*_{S^c}\|_1.\end{aligned}$$

Recall by assumption that

$$\kappa_2 - \tilde{\rho}_\gamma\kappa_1\frac{\|d_S\|_2}{d_{\min}} \geq \varepsilon,$$

then

$$\varepsilon\|\Delta\|_2 \leq \|\widetilde{A}\Delta\|_2 + \frac{\tilde{\rho}_\gamma\kappa_1}{d_{\min}}\|Dx^*_{S^c}\|_1.$$

Using the previous control of $\|\widetilde{A}\Delta\|_2^2$, we obtain

$$\begin{aligned}\|\Delta\|_2^2 \leq& \frac{2}{\varepsilon^2}\|\widetilde{A}\Delta\|_2^2 + \frac{2\tilde{\rho}_\gamma^2\kappa_1^2}{\varepsilon^2 d_{\min}^2}\|Dx^*_{S^c}\|_1^2\\ \leq& \frac{2\rho_\gamma}{\varepsilon^2}\|d_S\|_2\|\Delta\|_2 + \frac{2\rho_\gamma}{\varepsilon^2}\|Dx^*_{S^c}\|_1 + \frac{2\tilde{\rho}_\gamma^2\kappa_1^2}{\varepsilon^2 d_{\min}^2}\|Dx^*_{S^c}\|_1^2\\ \leq& \frac{2\rho_\gamma}{\varepsilon^2}\|d_S\|_2\|\Delta\|_2 + \sigma_S^2,\end{aligned}$$

where

$$\sigma_S^2 := \frac{2\rho_\gamma}{\varepsilon^2}\|Dx^*_{S^c}\|_1 + \frac{2\tilde{\rho}_\gamma^2\kappa_1^2}{\varepsilon^2 d_{\min}^2}\|Dx^*_{S^c}\|_1^2.$$

Recall $y^2 - by - c \leq 0$ implies $y^2 \leq b^2 + 2c$ for any y. Thus,

$$\begin{aligned}\|\Delta\|_2^2 \leq& \frac{4\rho_\gamma^2}{\varepsilon^4}\|d_S\|_2^2 + 2\sigma_S^2\\ \leq& \frac{4\rho_\gamma^2}{\varepsilon^4}\|d_S\|_2^2 + \frac{4\rho_\gamma}{\varepsilon^2}\|Dx^*_{S^c}\|_1 + \frac{4\tilde{\rho}_\gamma^2\kappa_1^2}{\varepsilon^2 d_{\min}^2}\|Dx^*_{S^c}\|_1^2.\end{aligned}$$

We conclude by observing that $\tilde{\rho}_\gamma \leq \rho_\gamma$.

□

B Concentration inequality for data-dependent weights (proof of Lemma 3.1)

The proof of (3.2) is classical and follows the lines of the proof of Bernstein's inequality. Let $Z = R^\top Y$ and $z = R^\top Ax^*$. Conditioned on the sensing matrix A, the Y_ℓ's are independent Poisson variables of mean $\sum_{k=1}^p a_{\ell,k}x_k^*$. Therefore for all $\lambda > 0$ (eventually depending only on the sensing matrix A)

$$\mathbb{E}\left(e^{\lambda(Z-z)}\big|A\right) = \prod_{\ell=1}^n \mathbb{E}\left(e^{\lambda R_\ell\left[Y_\ell - \sum_{k=1}^p a_{\ell,k}x_k^*\right]}\big|A\right) = \prod_{\ell=1}^n \exp\left[\left(e^{\lambda R_\ell} - \lambda R_\ell - 1\right)\sum_{k=1}^p a_{\ell,k}x_k^*\right].$$

If $\lambda < (3/b)$, then by classical computations (see [53] for instance), for all ℓ,

$$\left|e^{\lambda r_\ell} - \lambda r_\ell - 1\right| \leq \frac{\lambda^2 r_\ell^2/2}{1-\lambda b/3}.$$

Therefore, if $\lambda < (3/b)$,

$$\mathbb{E}\left(e^{\lambda(Z-z)}\big|A\right) \leq \exp\left(\frac{\lambda^2 v/2}{1-\lambda b/3}\right).$$

Hence by Markov's inequality, for all $u > 0$

$$\mathbb{P}(Z - z \geq u) \leq \exp\left(\frac{\lambda^2 v/2}{1 - \lambda b/3} - \lambda u\right).$$

It remains to optimize in λ and conclude as in Bernstein's inequality (see [53]). More precisely, the upper bound in λ is minimal if

$$\lambda = \frac{2\frac{bu}{3} + v - \sqrt{v^2 + 2\frac{buv}{3}}}{2\left(\frac{b^2 u}{9} + \frac{vb}{6}\right)},$$

which gives

$$\mathbb{P}(Z - z \geq u) \leq \exp\left(\frac{\sqrt{v^2 + 2\frac{buv}{3}} - v - \frac{bu}{3}}{b^2/9}\right).$$

We want the upper bound to be equal to $e^{-\theta}$, which gives by inversion

$$u = \sqrt{2v\theta} + \frac{b\theta}{3}.$$

For (3.3) it is sufficient to apply (3.2) to both R and $-R$. For (3.4) it is sufficient to apply (3.2) to $-\bar{R}_2$ and for (3.5), to combine both (3.3) and (3.4).

C Validation of assumptions for Bernoulli sensing of Section 4

C.1 Rescaling and recentering

First let us prove that $\mathbb{E}(\widetilde{G}) = I_p$, with $\widetilde{G} = \widetilde{A}^\top \widetilde{A}$. Indeed, the (k, k') element of $\widetilde{G}$ is

$$\widetilde{G}_{k,k'} = \frac{\sum_{\ell=1}^n (a_{\ell,k} - q)(a_{\ell,k'} - q)}{nq(1-q)}.$$

Hence $\mathbb{E}(\widetilde{G}_{k,k'}) = 0$ if $k \neq k'$ and $\mathbb{E}(\widetilde{G}_{k,k}) = 1$. Next let $Z = \widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)$ and let us prove that $\mathbb{E}(Z) = 0$.

$$\begin{aligned}
Z =& \frac{1}{nq(1-q)}(A^\top - q\mathbb{1}_{p\times 1}\mathbb{1}_{n\times 1}^\top)\left(\frac{nY - (\sum_{k=1}^p Y_k)\mathbb{1}_{n\times 1}}{n-1} - (A - q\mathbb{1}_{n\times 1}\mathbb{1}_{p\times 1}^\top)x^*\right) \\
=& \frac{1}{nq(1-q)}\left(\frac{n}{n-1}A^\top Y - \frac{\sum_{k=1}^p Y_k}{n-1}A^\top \mathbb{1}_{n\times 1} - A^\top A x^* + \right. \\
& \left. + q\|x^*\|_1 A^\top \mathbb{1}_{n\times 1} + q\mathbb{1}_{p\times 1}\mathbb{1}_{n\times 1}^\top A x^* - q^2\|x^*\|_1 \mathbb{1}_{p\times 1}\mathbb{1}_{n\times 1}^\top \mathbb{1}_{n\times 1}\right) \\
=& T_1 + T_2
\end{aligned}$$

with

$$T_1 = \frac{1}{nq(1-q)}\left(\frac{n}{n-1}A^\top - \frac{A^\top \mathbb{1}_{n\times 1}\mathbb{1}_{n\times 1}^\top}{n-1}\right)(Y - Ax^*)$$

and

$$T_2 = \frac{1}{nq(1-q)}\left(\frac{A^\top Ax^* - A^\top \mathbb{1}_{n\times 1}\mathbb{1}_{n\times 1}^\top Ax^*}{n-1} + q\|x^*\|_1 A^\top \mathbb{1}_{n\times 1} + q\mathbb{1}_{p\times 1}\mathbb{1}_{n\times 1}^\top Ax^* - q^2 n\|x^*\|_1\mathbb{1}_{p\times 1}\right).$$

Since $\mathbb{E}(Y|A) = Ax^*$, $\mathbb{E}(T_1|A) =$ and therefore $\mathbb{E}(T_1) = 0$.

Next the kth element of T_2 only depends on A and satisfies

$$\begin{aligned} nq(1-q)T_{2,k} =& \frac{1}{n-1}\left(\sum_{\ell=1}^{n}\sum_{k'=1}^{p} a_{\ell,k}a_{\ell,k'}x'_k - \sum_{\ell,\ell'=1}^{n}\sum_{k'=1}^{p} a_{\ell,k}a_{\ell',k'}x_{k'}\right) + q\sum_{\ell=1}^{n} a_{\ell,k}\sum_{k'=1}^{p} x_{k'} + \\ & q\sum_{\ell'=1}^{n}\sum_{k'=1}^{p} a_{\ell',k'}x_{k'} - q^2 n\sum_{k'=1}^{p} x_{k'} \\ =& \frac{-1}{n-1}\sum_{\ell=1}^{n}\sum_{\ell'\neq\ell}\sum_{k'=1}^{p}(a_{\ell,k}-q)(a_{\ell',k'}-q)x_{k'}. \end{aligned}$$

Hence every element of T_2 is a degenerate U-statistics of order 2 and $\mathbb{E}(T_2) = 0$. Note that T_2 can also be seen as $T_2 = \mathbb{M}x^*$ with

$$\mathbb{M}_{k,k'} = \frac{-1}{(n-1)nq(1-q)}\sum_{\ell=1}^{n}\sum_{\ell'\neq\ell}(a_{\ell,k}-q)(a_{\ell',k'}-q).$$

C.2 Assumption RE holds (proof of Proposition 4.1)

We use the following definition and lemma:

Definition C.1 (Zhou 2009, Definition 1.3)**.** *Let Y be a random vector in $\mathbb{R}^p$; Y is called isotropic if for every $y \in \mathbb{R}^p$, $\mathbb{E}|\langle Y, y\rangle|^2 = \|y\|_2^2$, and is $\psi_2(\alpha)$, for α a positive constant, if for every $y \in \mathbb{R}^p$*

$$\|\langle Y, y\rangle\|_{\psi_2} \leq \alpha\|y\|_2,$$

where for a random variable $X \in \mathbb{R}$

$$\|X\|_{\psi_2} := \inf\{t : \mathbb{E}\exp(X^2/t^2) \leq 2\}.$$

Lemma C.1 (Theorem 3 in Li and Raskutti, Minimax optimal convex methods for Poisson inverse problems under ℓ_q-ball sparsity)**.** *There exist positive constants c, c', c'' for which the following holds. Let B be an isotropic $\psi_2(\alpha)$ random vector of $\mathbb{R}^p$. Let $B_1, \ldots, B_n \in \mathbb{R}^p$ be independent, distributed according to B and define $\Gamma = \sum_{i=1}^n \langle B_i, \cdot\rangle e_i$, where e_i is the ith element of the standard basis of $\mathbb{R}^n$. Then with probability at least $1 - c'\exp(-c''n)$, for all $x \in \mathbb{R}^p$ we have*

$$\frac{\|x\|_2}{4} - c\alpha^2\sqrt{\frac{\log p}{n}}\|x\|_1 \leq \frac{\|\Gamma x\|_2}{\sqrt{n}}.$$

We apply this lemma with

$$\Gamma = \sqrt{n} \times \widetilde{A}.$$

Therefore, for any i and any j,

$$(B_i)_j = \frac{A_{ij} - q}{\sqrt{q(1-q)}}.$$

and we have for any $x \in \mathbb{R}^p$ and any i,

$$\mathbb{E}\left[\langle B_i, x\rangle^2\right] = \operatorname{Var}\left(\sum_{j=1}^{p} (B_i)_j x_j\right) = \|x\|_2^2,$$

so the B_i's are isotropic. Furthermore, using Lemma 5.9 of [54]

$$\|\langle B_i, x\rangle\|_{\psi_2}^2 \le \Box \sum_{j=1}^{p} \|x_j (B_i)_j\|_{\psi_2}^2 \le \Box \sum_{j=1}^{p} x_j^2 \|(B_i)_j\|_{\psi_2}^2.$$

Since

$$|(B_i)_j| \le \frac{1}{\sqrt{q(1-q)}},$$

using Lemma 5.5 and Example 5.8 of [54], we obtain

$$\|(B_i)_j\|_{\psi_2} \le \frac{1}{\sqrt{q(1-q)}}$$

and

$$\|\langle B_i, x\rangle\|_{\psi_2}^2 \le \Box \times \frac{\|x\|_2^2}{q(1-q)}.$$

So, in our setting, the B_i's are $\psi_2(\alpha)$ with

$$\alpha = \frac{1}{\sqrt{q(1-q)}}.$$

So, with probability at least $1 - c' \exp(-c'' n)$, for all $x \in \mathbb{R}^p$,

$$\|\widetilde{A} x\|_2 = \frac{\|\Gamma x\|_2}{\sqrt{n}} \ge \frac{\|x\|_2}{4} - \frac{c}{q(1-q)} \sqrt{\frac{\log p}{n}} \|x\|_1,$$

and Assumption $RE(\kappa_1, \kappa_2)$ holds with $\kappa_2 = \frac{1}{4}$ and $\kappa_1 = \frac{c}{q(1-q)} \sqrt{\frac{\log p}{n}}$.

C.3 Proofs for data-dependent weights

Proposition C.1. *Let*

$$W = \max_{u,k=1,\dots,p} w(u,k)$$

with

$$w(u,k) = \frac{1}{n^2 (n-1)^2 q^2 (1-q)^2} \sum_{\ell=1}^{n} a_{\ell,u} \left(n a_{\ell,k} - \sum_{\ell'=1}^{n} a_{\ell',k} \right)^2.$$

Then if $nq \ge 12 \max(q, 1-q) \log(p)$, and if $p \ge 2$ then there exists absolute constants c, c' such that with probability larger than $1 - c'/p$, the choice

$$d = \sqrt{6 W \log(p)} \sqrt{\hat{N}} + \frac{\log(p)}{(n-1) q (1-q)} + c \left(\frac{3 \log(p)}{n} + \frac{9 \max(q^2, (1-q)^2)}{n^2 q (1-q)} \log(p)^2 \right) \hat{N}$$

satisfies Assumption Weights(d), where $\hat{N}$ is an estimator of $\|x^*\|_1$ given by

$$\hat{N} = \frac{1}{nq - \sqrt{6nq(1-q)\log(p)} - \max(q, 1-q)\log(p)} \left(\sqrt{\frac{3\log(p)}{2}} + \sqrt{\frac{5\log(p)}{2} + \sum_{\ell=1}^{n} Y_\ell} \right)^2 .$$

One can take $c = 126$ as long as $n \geq 20$.

Proposition C.2. *There exists some absolute constant κ such that if*

$$nq^2(1-q) \geq \kappa \log(p),$$

then there exists a positive constant C such that with probability larger than $1 - C/p$,

$$d \simeq \sqrt{\frac{\log(p)\|x^*\|_1}{n\min(q, 1-q)}} + \frac{\log(p)\|x^*\|_1}{n} + \frac{\log(p)}{nq(1-q)}.$$

Proposition C.3. *With the same notations and assumptions as Proposition C.1, there exists absolute constants c, c' such that with probability larger than $1 - c'/p$, the choice (depending on k)*

$$d_k = \sqrt{6\log(p)} \left(\sqrt{\frac{3\log(p)}{2(n-1)^2q^2(1-q)^2}} + \sqrt{\frac{5\log(p)}{2(n-1)^2q^2(1-q)^2} + V_k^\top Y} \right) + \frac{\log(p)}{(n-1)q(1-q)} + \tag{C.1}$$

$$+c\left(\frac{3\log(p)}{n} + \frac{9\max(q^2, (1-q)^2)}{n^2q(1-q)}\log(p)^2\right)\hat{N}, \tag{C.2}$$

satisfies Assumption Weights($\{d_k\}_k$), where the vector V_k of size n is given by

$$V_{k,\ell} = \left(\frac{na_{\ell,k} - \sum_{\ell'=1}^{n} a_{\ell',k}}{n(n-1)q(1-q)} \right)^2 .$$

One can take $c = 126$ as long as $n \geq 20$.

Proposition C.4. *There exists some absolute constant κ such that if*

$$nq^2(1-q) \geq \kappa \log(p),$$

then there exists a positive C such that with probability larger than $1 - C/p$

$$d_k \simeq \sqrt{\log(p)\left[\frac{x_k^*}{nq} + \frac{\sum_{u\neq k} x_u^*}{n(1-q)}\right]} + \frac{\log(p)\|x^*\|_1}{n} + \frac{\log(p)}{nq(1-q)}.$$

C.3.1 Assumption Weights holds (proof of Propositions C.1 and C.3)

As shown in Appendix C.1,

$$(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k \leq T_{1,k} + T_{2,k}.$$

To derive the constant weight of Proposition C.1, we use a bound on $\|T_1\|_\infty + \|T_2\|_\infty$. To derive the non-constant weights of Proposition C.3, we use a bound on $|T_{1,k}| + \|T_2\|_\infty$. These bounds are derived in this section.

Concentration of T_2 Each element of the matrix $\mathbb{M}$ is a degenerate U-statistics of order 2 of the form $2U$ with $U = \sum_{\ell > \ell'} g(a_{\ell,k}, a_{\ell',k'})$ to which one can apply [55]. Let us compute the different quantities involved in this concentration formula.

Since $q(1-q) \leq (q^2 + (1-q)^2)/2 \leq \max(q^2, (1-q)^2)$, a deterministic upper bound of g does not depend on k, k' and is given by

$$A_{\mathbb{M}} = \frac{\max(q^2, (1-q)^2)}{n(n-1)q(1-q)}.$$

On the other hand for any $a \in \{0, 1\}$,

$$\mathbb{E}(g^2(a_{\ell,k}, a)) = \frac{(a-q)^2}{n^2(n-1)^2 q(1-q)}.$$

Therefore $C_{\mathbb{M}}^2 = \frac{1}{2n(n-1)}$ and

$$B_{\mathbb{M}}^2 = \frac{\max(q^2, (1-q)^2)}{n^2(n-1)q(1-q)}.$$

Finally $D_{\mathbb{M}}$ should be chosen as an upper bound of

$$\mathbb{E}\left(\sum_{\ell > \ell'} g(a_{\ell,k}, a_{\ell',k'}) c_\ell(a_{\ell,k}) b_{\ell'}(a_{\ell',k'})\right),$$

for all choices of functions c_ℓ, $b_{\ell'}$ such that $\mathbb{E}(\sum_{\ell=2}^n c_\ell(a_{\ell,k})^2) \leq 1$ and $\mathbb{E}(\sum_{\ell'=1}^{n-1} b_{\ell'}(a_{\ell',k'})^2) \leq 1$. Note

$$\sum_{\ell'=1}^{\ell} \mathbb{E}\left((a_{\ell',k'} - q) b_{\ell'}(a_{\ell',k'})\right) \leq \sqrt{\sum_{\ell'} \mathbb{E}(((a_{\ell',k'} - q)^2)} \sqrt{\sum_{\ell'} \mathbb{E}(b_{\ell'}(a_{\ell',k'})^2)} \leq \sqrt{nq(1-q)}.$$

By doing the same for the terms in $a_{\ell,k}$, $D_{\mathbb{M}} = \frac{nq(1-q)}{n(n-1)q(1-q)} = \frac{1}{n-1}$ works. By Theorem 3.4.8 of [56], for all $\theta > 0$, with probability larger than $1 - c_1 p^2 e^{-\theta}$, for some absolute constant c_1 ($c_1 = 5.4$ works), we have that for all k, k'

$$\begin{aligned} |\mathbb{M}_{k,k'}| &\leq 8C_{\mathbb{M}}\sqrt{\theta} + 8\sqrt{2}D_{\mathbb{M}}\theta + 216(B_{\mathbb{M}}\theta^{3/2} + A_{\mathbb{M}}\theta^2) \\ &\leq \frac{8\sqrt{\theta}}{\sqrt{2n(n-1)}} + \frac{8\sqrt{2}\theta}{(n-1)} + 216\left(\sqrt{\frac{\theta}{n}\frac{\max(q^2, (1-q)^2)\theta^2}{n(n-1)q(1-q)}} + \frac{\max(q^2, (1-q)^2)\theta^2}{n(n-1)q(1-q)}\right). \end{aligned}$$

Let us take $\theta > 1$ so that $\sqrt{\theta} \leq \theta$. By using the classical inequality $ab \leq (a^2 + b^2)/2$, we end up with

$$|\mathbb{M}_{k,k'}| \leq \left(\frac{8}{\sqrt{2}} + 8\sqrt{2}\right)\frac{\theta}{n-1} + 108\frac{\theta}{n} + 108\frac{\max(q^2, (1-q)^2)\theta^2}{n(n-1)q(1-q)}.$$

Therefore there exists some absolute constants c_1 and c_2 such that as soon as $n \geq 2$ and $\theta > 1$, with probability larger than $1 - c_1 p^2 e^{-\theta}$, for all k, k'

$$|\mathbb{M}_{k,k'}| \leq c_2\left(\frac{\theta}{n} + \frac{\max(q^2, (1-q)^2)}{n^2 q(1-q)}\theta^2\right), \tag{C.3}$$

where $c_2 = 126$ works as soon as $n \geq 20$. Therefore on the same event,

$$\|T_2\|_\infty \leq c_2\left(\frac{\theta}{n} + \frac{\max(q^2, (1-q)^2)}{n^2 q(1-q)}\theta^2\right)\|x^*\|_1. \tag{C.4}$$

Concentration around $\|x^*\|_1$ Since $\|x^*\|_1$ is unknown in the previous inequality, if one wants to upper bound T_2, we need to estimate it.

Applying (3.4) of Lemma 3.1 with $R = \mathbb{1}_{n\times 1}$ gives that with probability larger than $1-e^{-\theta}$,

$$\bar{x}_a = \sum_{\ell,k} a_{\ell,k} x_k^* \leq \left(\sqrt{\frac{\theta}{2}} + \sqrt{\frac{5\theta}{6} + \sum_{\ell=1}^{n} Y_\ell}\right)^2.$$

But by using Bernstein's inequality, with probability larger than $1-2pe^{-\theta}$, for all k,

$$|\sum_{\ell=1}^{n}(a_{\ell,k}-q)| \leq C_{n,\theta} = \sqrt{2nq(1-q)\theta} + \max(q,(1-q))\frac{\theta}{3}. \tag{C.5}$$

Hence on this event,

$$(nq - C_{n,\theta})\|x\|_1 \leq \bar{x}_a.$$

So the first assumption on the range of (n,q) is to assume that $nq > C_{n,\theta}$, which is implied by

$$nq \geq 4\max(q, 1-q)\theta. \tag{C.6}$$

In this case, with probability larger than $1-(2p+1)e^{-\theta}$,

$$\|x\|_1 \leq \hat{N}_\theta := \frac{1}{nq - C_{n,\theta}}\left(\sqrt{\frac{\theta}{2}} + \sqrt{\frac{5\theta}{6} + \sum_{\ell=1}^{n} Y_\ell}\right)^2. \tag{C.7}$$

Hence there exists some absolute constant c_3 such that on an event of probability larger than $1-c_3p^2e^{-\theta}$,

$$\|T_2\|_\infty \leq c_2\left(\frac{\theta}{n} + \frac{\max(q^2,(1-q)^2)}{n^2q(1-q)}\theta^2\right)\hat{N}_\theta. \tag{C.8}$$

Upper-bound for T_1 The upper bound on T_2 does not depend on k, but it is just a residual term. The upper bound for T_1 gives the main tendency and its behavior may be refined k by k leading to a weight d_k that depends on k. Recall that for fixed k, $T_{1,k} = R_k^\top(Y - Ax^*)$ with for all $\ell = 1,\dots,n$,

$$R_{k,\ell} = \frac{na_{\ell,k} - \sum_{\ell'=1}^{n} a_{\ell',k}}{n(n-1)q(1-q)}.$$

By (3.3) of Lemma 3.1, on an event of probability larger than $1-2pe^{-\theta}$,

$$|T_{1,k}| \leq \sqrt{2V_k^\top Ax^*\theta} + \frac{\|R_k\|_\infty\theta}{3},$$

with $V_{k,\ell} = R_{k,\ell}^2$. But since the $a_{\ell,k}$ have values in $\{0,1\}$, one has that

$$\|R_k\|_\infty \leq \frac{1}{(n-1)q(1-q)}.$$

Moreover,

$$V_k^\top Ax^* \leq W\|x^*\|_1,$$

with

$$W = \max_{u,k=1,\dots,p} w(u,k)$$

and

$$w(u,k) = \frac{1}{n^2(n-1)^2q^2(1-q)^2}\sum_{\ell=1}^{n} a_{\ell,u}\left(na_{\ell,k} - \sum_{\ell'=1}^{n} a_{\ell',k}\right)^2.$$

Combined with (C.7), this gives that

$$\|T_1\|_\infty \leq \sqrt{2W\theta}\sqrt{\hat{N}} + \frac{\theta}{3(n-1)q(1-q)}. \tag{C.9}$$

This combined with (C.8) and the choice $\theta = 3\log(p)$ (which is larger than 1 since $p \geq 2$) gives Proposition C.1. On the other hand, one could have applied (3.5) of Lemma 3.1 to obtain that on an event of probability larger than $1 - 3pe^{-\theta}$,

$$|T_{1,k}| \leq \left(\sqrt{\frac{\theta}{2(n-1)^2q^2(1-q)^2}} + \sqrt{\frac{5\theta}{6(n-1)^2q^2(1-q)^2} + V_k^\top Y}\right)\sqrt{2\theta} + \frac{\theta}{3(n-1)q(1-q)}.$$

Once again, this combined with (C.8) and the choice $\theta = 3\log(p)$ gives Proposition C.3.

C.3.2 Bounds on the $w(u,k)$'s

First let us remark that if we denote

$$w_1(u,k) = \frac{1}{(n-1)^2q^2(1-q)^2}\sum_{\ell=1}^{n} a_{\ell,u}(a_{\ell,k} - q)^2$$

and

$$w_2(u,k) = \frac{1}{n^2(n-1)^2q^2(1-q)^2}\left(\sum_{\ell=1}^{n} a_{\ell,u}\right)\left(\sum_{\ell'=1}^{n}(a_{\ell,k} - q)\right)^2.$$

Then for all $\epsilon \in (0,1)$,

$$(1-\epsilon)w_1(u,k) + (1-\frac{1}{\epsilon})w_2(u,k) \leq w(u,k) \leq (1+\epsilon)w_1(u,k) + (1+\frac{1}{\epsilon})w_2(u,k).$$

In the sequel we consequently need to find an upper-bound for $w_2(u,k)$ and a lower and upper bound on $w_1(u,k)$ to obtain bounds for $w(u,k)$.

Upper bound for $w_2(u,k)$ By (C.5) and remarking that $\max(q, 1-q) \leq 1$, on an event of probability larger than $1 - 2pe^{-\theta}$,

$$\begin{aligned}
w_2(u,k) \leq & \frac{nq + \sqrt{2nq(1-q)\theta} + \frac{\theta}{3}}{n^2(n-1)^2q^2(1-q)^2}\left(\sqrt{2nq(1-q)\theta} + \frac{\theta}{3}\right)^2 \\
\leq\square & \frac{n^2q^2(1-q)\theta + nq\theta^2 + (nq(1-q)\theta)^{3/2} + \theta^3}{n^4q^2(1-q)^2} \\
\leq\square & \left(\frac{\theta}{n^2(1-q)} + \frac{\theta^2}{n^3q(1-q)^2} + \frac{\theta^{3/2}}{n^{5/2}q^{1/2}(1-q)^{1/2}} + \frac{\theta^3}{n^4q^2(1-q)^2}\right).
\end{aligned}$$

If one assumes that

$$nq(1-q) \geq \theta, \tag{C.10}$$

then the leading term in the previous expansion is the first one and

$$w_2(u,k) \leq \square \frac{\theta}{n^2(1-q)}. \tag{C.11}$$

Now for the control of $w_1(u,k)$, if $u=k$ then on can rewrite

$$\begin{aligned}
w_1(k,k) =& \frac{1}{(n-1)^2q^2(1-q)^2}\sum_{\ell=1}^n a_{\ell,k}(a_{\ell,k}-q)^2\\
=& \frac{1}{(n-1)^2q^2(1-q)^2}\sum_{\ell=1}^n (a_{\ell,k}^3 - 2qa_{\ell,k}^2 + q^2 a_{\ell,k})\\
=& \frac{1}{(n-1)^2q^2(1-q)^2}\sum_{\ell=1}^n a_{\ell,k}(1-q)^2\\
=& \frac{1}{(n-1)^2q^2}\sum_{\ell=1}^n a_{\ell,k}.
\end{aligned}$$

So by (C.5), on the same event as before, because of (C.10),

$$\left| w_1(k,k) - \frac{n}{(n-1)^2q}\right| \leq \frac{\sqrt{2nq(1-q)\theta}+\frac{\theta}{3}}{(n-1)^2q^2} \leq \square \frac{(1-q)^{1/2}\theta^{1/2}}{n^{3/2}q^{3/2}}. \tag{C.12}$$

On the other hand, if $u \neq k$, let us apply Bernstein inequality to $Z_\ell^{u,k} = a_{\ell,u}(a_{\ell,k}-q)^2$. The expectation of $Z_\ell^{u,k}$ is given by

$$\mathbb{E}(Z_\ell^{u,k}) = q^2(1-q),$$

whereas its variance is

$$\begin{aligned}
\mathrm{Var}(Z_\ell^{u,k}) =& \mathbb{E}([Z_\ell^{u,k}]^2) - q^4(1-q)^2\\
=& \mathbb{E}(a_{\ell,u}^2)\mathbb{E}([a_{\ell,k}-q]^4) - q^4(1-q)^2\\
=& q\mathbb{E}(a_{\ell,k}^4 - 4qa_{\ell,k}^3 + 6q^2a_{\ell,k}^2 - 4q^3a_{\ell,k} + q^4) - q^4(1-q)^2\\
=& q(q-4q^2+6q^3-3q^4) - q^4(1-q)^2\\
=& q^2(1-q)(1-3q+3q^2) - q^4(1-q)^2\\
=& q^2(1-q)(1-3q+2q^2+q^3)\\
\leq& q^2(1-q).
\end{aligned}$$

Moreover $|Z_\ell^{u,k}|$ is bounded by 1. So Bernstein inequality gives that with probability larger than $1-2p(p-1)e^{-\theta}$,

$$\left|\sum_{\ell=1}^n Z_\ell^{u,k} - nq^2(1-q)\right| \leq \sqrt{2nq^2(1-q)\theta} + \frac{\theta}{3}. \tag{C.13}$$

Hence on the same event, because of (C.10), if we additionally assume that

$$nq^2(1-q) \geq \theta, \tag{C.14}$$

we have

$$\left|w_1(u,k) - \frac{n}{(n-1)^2(1-q)}\right| \leq \frac{\sqrt{2nq^2(1-q)\theta} + \frac{\theta}{3}}{(n-1)^2q^2(1-q)^2} \leq \square \frac{\theta^{1/2}}{n^{3/2}q(1-q)^{3/2}}. \tag{C.15}$$

So finally there is a constant $\kappa(\epsilon)$ such that if

$$nq^2(1-q) \geq \kappa(\epsilon)\theta, \tag{C.16}$$

then on this event of probability larger than $1 - \square p^2 e^{-\theta}$,

$$(1-\epsilon)\frac{1}{nq} \leq w_1(k,k) \leq (1+\epsilon)\frac{1}{nq},$$

and if $u \neq k$,

$$(1-\epsilon)\frac{1}{n(1-q)} \leq w_1(u,k) \leq (1+\epsilon)\frac{1}{n(1-q)}.$$

Hence since (C.11) holds, on the same event,

$$(1-\epsilon)^2\frac{1}{nq} + (1-\frac{1}{\epsilon})\square\frac{\theta}{n^2(1-q)} \leq w(k,k) \leq (1+\epsilon)^2\frac{1}{nq} + (1+\frac{1}{\epsilon})\square\frac{\theta}{n^2(1-q)}.$$

This implies up to the eventual replacement of $\kappa(\epsilon)$ by a bigger constant still depending on ϵ that

$$(1-\epsilon)^3\frac{1}{nq} \leq w(k,k) \leq (1+\epsilon)^3\frac{1}{nq}, \tag{C.17}$$

and in the same way that for $u \neq k$ that

$$(1-\epsilon)^3\frac{1}{n(1-q)} \leq w(u,k) \leq (1+\epsilon)^3\frac{1}{n(1-q)}. \tag{C.18}$$

C.3.3 Control of the constant weight (proof of Proposition C.2)

Applying (3.2) of Lemma 3.1 with $R = \mathbb{1}_{n\times 1}$ gives that with probability larger than $1 - e^{-\theta}$,

$$\sum_{\ell=1}^{n} Y_\ell \leq \square\left(\sum_{\ell,k} a_{\ell,k}x_k^* + \theta\right).$$

Then by using (C.5), we get that on an event of probability larger than $1 - \square pe^{-\theta}$

$$\sum_{\ell=1}^{n} Y_\ell \leq \square\left((nq + C_{n,\theta})\|x^*\|_1 + \theta\right).$$

This implies that on the same event

$$\hat{N} \leq \square\frac{nq + C_{n,\theta}}{nq - C_{n,\theta}}\|x^*\|_1 + \square\frac{\theta}{nq - C_{n,\theta}}.$$

By eventually increasing $\kappa(\epsilon)$ again, we have that under (C.16)

$$\hat{N} \leq \square\frac{1+\epsilon}{1-\epsilon}\|x^*\|_1 + \square\frac{\theta}{(1-\epsilon)nq}.$$

Hence, combining with (C.7), on an event of probability larger that $1 - \square p e^{-\theta}$,

$$\|x^*\|_1 \leq \hat{N} \leq \square \frac{1+\epsilon}{1-\epsilon}\|x^*\|_1 + \square \frac{\theta}{(1-\epsilon)nq}.$$

Hence, using again (C.16), with eventually a larger κ and fixing $\epsilon = 1/2$ say, gives

$$\square\left[\sqrt{W\theta\|x^*\|_1} + \frac{\theta\|x^*\|_1}{n} + \frac{\theta}{nq(1-q)}\right] \leq d \leq \square\left[\sqrt{W\theta\left(\|x^*\|_1 + \frac{\theta}{nq}\right)} + \frac{\theta\|x^*\|_1}{n} + \frac{\theta}{nq(1-q)}\right].$$

By previous computations, W is of the order of $\frac{1}{n\min(q,1-q)}$, which gives Proposition C.2 with $\theta = 3\log(p)$.

C.3.4 Control of the non-constant weights (proof of Proposition C.4)

Similarly, applying (3.2) and (3.4) of Lemma 3.1 to $V_k^\top Y$ with V_k for all k gives that with probability larger than $1 - \square p e^{-\theta}$, for all k,

$$V_k^\top A x^* \leq \left(\sqrt{\frac{\theta}{2(n-1)^2 q^2(1-q)^2}} + \sqrt{\frac{5\theta}{6(n-1)^2 > q^2(1-q)^2} + V_k^\top Y}\right)^2 \leq \square\left(V_k^\top A x^* + \frac{\theta}{n^2q^2(1-q)^2}\right).$$

But $V_k^\top A x^* = \sum_{u=1}^p w(u,k) x_u^*$ which is of the order of

$$\frac{1}{nq} x_k^* + \frac{1}{n(1-q)} \sum_{u \neq k} x_u^*.$$

This gives Proposition C.4 with $\theta = 3\log(p)$.

C.4 Proof of Proposition 4.4

Proof of Proposition 4.4. By denoting $\widetilde{A}_{S^*}$ the matrix of size $n \times |S^*|$ whose columns are the columns of $\widetilde{A}$ corresponding to non-zero elements of x^*, we have for any $k \in S^*$,

$$\widehat{x}_k^{\mathrm{OLS}} = ((\widetilde{A}_{S^*}^H \widetilde{A}_{S^*})^{-1} \widetilde{A}_{S^*}^H \widetilde{Y})_k = (\widetilde{G}_{S^*}^{-1} \widetilde{A}_{S^*}^H \widetilde{Y})_k,$$

where $\widetilde{G}_{S^*} = \widetilde{A}_{S^*}^H \widetilde{A}_{S^*}$. Therefore, by setting $\widehat{x}_k^{\mathrm{OLS}} = 0$ for $k \notin S^*$, we have

$$\|\widehat{x}^{\mathrm{OLS}} - x^*\|_2^2 = \|\widetilde{G}_{S^*}^{-1} \widetilde{A}_{S^*}^H (\widetilde{Y} - \widetilde{A} x^*)\|_2^2.$$

Theorem 2.4 in [57] shows that there exist constants $c_1, c_2, c_3, C > 0$ such that for any $\delta_{s^*} \in (0, 1/2)$, if

$$s^* \leq \frac{c_1 \delta_{s^*}^2 n}{\alpha_q^4 \log(c_2 p \alpha_q^4 / \delta_{s^*}^2 n)}, \qquad \text{and} \qquad n \geq \frac{\alpha_q^4}{c_3 \delta_{s^*}^2} \log p, \tag{C.19}$$

where

$$\alpha_q := \begin{cases} \sqrt{\frac{3}{2q(1-q)}}, & q \neq 1/2 \\ 1, & q = 1/2 \end{cases},$$

then for any s^*-sparse x,

$$(1 - \delta_{s^*})\|x\|_2^2 \leq \|\widetilde{A} x\|_2^2 \leq (1 + \delta_{s^*})\|x\|_2^2$$

with probability exceeding $1 - C/p$ for a universal positive constant C. Now, assume that (4.8) is satisfied. Then, (C.19) is satisfied for $\delta_{s^*} = 1/4$. Therefore, all eigenvalues of $\widetilde{G}_{S^*}$ are included in the interval $[3/4\ ;\ 5/4]$ and we obtain the result. □

D Validation of assumptions for random convolution model of Section 5

D.1 Rescaling and recentering

Note that Proposition D.1 given in the next section proves in particular that $\mathbb{E}(\widetilde{G}) = I_p$. By Lemma D.1 below, we obtain in particular that $\mathbb{E}(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)) = 0$ as expected.

Lemma D.1. *Conditionally on the U_i's, $\widetilde{Y}$ is an unbiased estimate of $\widetilde{A}x^*$:*

$$\mathbb{E}[\widetilde{Y}|U_1, \dots, U_m] = \widetilde{A}x^*.$$

Proof of Lemma D.1. We have first

$$\begin{aligned}\mathbb{E}[\overline{Y}] =& \frac{1}{m}\sum_{\ell=1}^{p}\mathbb{E}[(Ax^*)_\ell] = \frac{1}{m}\sum_{\ell=1}^{p}\sum_{k=1}^{p}\mathbb{E}[a_{\ell,k}x_k^*] \\ =& \frac{1}{m}\sum_{\ell=1}^{p}\sum_{k=1}^{p}x_k^*\mathbb{E}[\mathbb{N}(\ell-k)] = \frac{1}{m}\sum_{k=1}^{p}x_k^* m = \|x^*\|_1.\end{aligned}$$

The result can be now deduced:

$$\begin{aligned}\mathbb{E}[\widetilde{Y}|U_1, \dots, U_m] =& \frac{1}{\sqrt{m}}\left[Ax^* - \frac{m-\sqrt{m}}{p}\sum_{\ell=1}^{p}x_\ell^*\mathbb{1}\right] \\ =& \frac{1}{\sqrt{m}}\left[A - \frac{m-\sqrt{m}}{p}\mathbb{1}\mathbb{1}^\top\right]x^* \\ =& \widetilde{A}x^*.\end{aligned}$$

□

D.2 Assumption RE holds (proof of Proposition 5.1)

Proof of Proposition 5.1. For this purpose, let us introduce the following degenerate U-statistics of order two, defined for all $k \in \{0, \dots, p-1\}$ by

$$\mathbb{U}(k) = \sum_{u=1}^{p}\sum_{i=1}^{m}\sum_{j\neq i, j=1}^{m}\left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\left(\mathbf{1}_{U_j=u+k[p]} - \frac{1}{p}\right). \tag{D.1}$$

Proposition D.1. *Let $p, m > 1$ be fixed integers. For any $k, \ell \in \{0, \dots, p-1\}$, we have:*

$$(\widetilde{G} - I_p)_{k,\ell} = \frac{1}{m}\mathbb{U}(k-\ell).$$

Furthermore, there exists absolute positive constants κ such that for all real number $\theta > 1$ such that there exists an event $\Omega_{\mathbb{U}}(\theta)$ of probability larger than $1 - 5.54\ pe^{-\theta}$ and, on this event, for all $k \in \{0, \dots, p-1\}$,

$$|\mathbb{U}(k)| \leq m\xi(\theta) \tag{D.2}$$

with

$$\xi(\theta) = \kappa\left(\frac{\theta}{\sqrt{p}} + \frac{\theta^2}{m}\right).$$

Note that this proves that the first part of Proposition 5.1 is satisfied on the event $\Omega_{\mathbb{U}}(\theta)$ with $\theta = 2\log p$. On this event,

$$\|\widetilde{G} - I_p\|_\infty \leq \xi,$$

where

$$\widetilde{G} := \widetilde{A}^\top \widetilde{A}$$

and

$$\xi = \kappa \left(\frac{\log p}{\sqrt{p}} + \frac{\log^2 p}{m} \right).$$

Thus for any $x \in \mathbb{R}^p$ we have

$$\begin{aligned} \|x\|_2^2 =& x^\top (\widetilde{G} - \widetilde{G} + I_p)x \\ =& \|\widetilde{A}x\|_2^2 + x^\top (I_p - \widetilde{G})x \\ \leq& \|\widetilde{A}x\|_2^2 + \xi\|x\|_1^2, \\ \|x\|_2 \leq& \|\widetilde{A}x\|_2 + \sqrt{\xi}\|x\|_1, \end{aligned}$$

and Assumption $RE(\kappa_1, \kappa_2)$ holds with $\kappa_2 = 1$ and $\kappa_1 = \sqrt{\xi}$. □

Proof of Proposition D.1. Let $\beta_0 = \frac{1}{\sqrt{m}}$ and $\beta_1 = \frac{\sqrt{m}-1}{p}$. For all $k \neq \ell \in \{0, \ldots, p-1\}$,

$$(A^\top A)_{k,k} = \sum_{u=1}^{p} \mathbb{N}(u)^2, \tag{D.3a}$$

$$(A^\top A)_{k,\ell} = \sum_{u=1}^{p} \mathbb{N}(u)\mathbb{N}(u + k - \ell). \tag{D.3b}$$

First note that

$$\begin{aligned} \mathbb{U}(d) =& \sum_u \sum_{i \neq j} \mathbf{1}_{U_i=u}\mathbf{1}_{U_j=u+d} - \frac{m-1}{p}\sum_{j=1}^{m}\sum_u \mathbf{1}_{U_j=u+d} - \frac{m-1}{p}\sum_{i=1}^{m}\sum_u \mathbf{1}_{U_i=u} + \frac{m(m-1)p}{p^2} \\ =& \sum_u \sum_{i \neq j} \mathbf{1}_{U_i=u}\mathbf{1}_{U_j=u+d} - \frac{m(m-1)}{p}. \end{aligned}$$

If $d \neq 0$,

$$\sum_u \sum_{i \neq j} \mathbf{1}_{U_i=u}\mathbf{1}_{U_j=u+d} = \sum_u \sum_{i,j} \mathbf{1}_{U_i=u}\mathbf{1}_{U_j=u+d} = \sum_u \mathbb{N}(u)\mathbb{N}(u+d),$$

and

$$\mathbb{U}(d) = \sum_u \mathbb{N}(u)\mathbb{N}(u+d) - \frac{m(m-1)}{p}. \tag{D.4}$$

If $d = 0$, if $U_i = u$ then $\sum_{j \neq i} \mathbf{1}_{U_j=u} = \mathbb{N}(u) - 1$ and

$$\sum_{i \neq j} \mathbf{1}_{U_i=u}\mathbf{1}_{U_j=u+d} = \mathbb{N}(u)(\mathbb{N}(u) - 1),$$

which leads to

$$\mathbb{U}(0) = \sum_u \mathbb{N}(u)(\mathbb{N}(u) - 1) - \frac{m(m-1)}{p} = \sum_u \mathbb{N}(u)^2 - m - \frac{m(m-1)}{p}. \tag{D.5}$$

Thus

$$(A^\top A)_{k,\ell} = \begin{cases} \mathbb{U}(0) + m + \frac{m(m-1)}{p}, & \text{if } k = \ell, \\ \mathbb{U}(k-\ell) + \frac{m(m-1)}{p}, & \text{if } k \neq \ell. \end{cases}$$

Next note that

$$\begin{aligned} \widetilde{G} =& \widetilde{A}^\top \widetilde{A} = (\beta_0 A - \beta_1 \mathbb{1}\mathbb{1}^\top)^\top (\beta_0 A - \beta_1 \mathbb{1}\mathbb{1}^\top) && \text{(D.6)} \\ =& \beta_0^2 A^\top A - \beta_0\beta_1(\mathbb{1}\mathbb{1}^\top A + A^\top \mathbb{1}\mathbb{1}^\top) + \beta_1^2 p \mathbb{1}\mathbb{1}^\top, && \text{(D.7)} \\ \widetilde{G}_{k,\ell} =& \beta_0^2 (A^\top A)_{k,\ell} - 2\beta_0\beta_1 m + \beta_1^2 p. && \text{(D.8)} \end{aligned}$$

For $k \neq \ell$, we have

$$\begin{aligned} \widetilde{G}_{k,\ell} =& \beta_0^2(\mathbb{U}(k-\ell) + \frac{m(m-1)}{p}) - 2\beta_0\beta_1 m + \beta_1^2 p \\ =& \frac{1}{m}(\mathbb{U}(k-\ell) + \frac{m(m-1)}{p}) - 2\sqrt{m}\frac{\sqrt{m}-1}{p} + \frac{(\sqrt{m}-1)^2}{p} \\ =& \frac{1}{m}\mathbb{U}(k-\ell). \end{aligned}$$

Similarly, for $k = \ell$, we have

$$\begin{aligned} \widetilde{G}_{k,k} =& \beta_0^2(\mathbb{U}(0) + m + \frac{m(m-1)}{p}) - 2\beta_0\beta_1 m + \beta_1^2 p \\ =& \frac{1}{m}\mathbb{U}(k-\ell) + 1. \end{aligned}$$

For the second result, one can rewrite $\mathbb{U}(d)$ as $\mathbb{U}(d) = \sum_{i<j} g(U_i, U_j)$, with

$$g(U_i, U_j) = \sum_{u=1}^{p} \left\{ \left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\left(\mathbf{1}_{U_j=u+d} - \frac{1}{p}\right) + \left(\mathbf{1}_{U_i=u+d} - \frac{1}{p}\right)\left(\mathbf{1}_{U_j=u} - \frac{1}{p}\right) \right\}.$$

Therefore $\mathbb{U}(d)$ is a completely degenerate U-statistic of order 2, and one can apply concentration inequalities of [55]. One can identify the corresponding constants $A_\mathbb{U}, B_\mathbb{U}, C_\mathbb{U}, D_\mathbb{U}$ as follows. The constant $A_\mathbb{U}$ should be an upper bound of $\|g\|_\infty$ but for $a, b \in \{0, \dots, p-1\}$, the largest value for $|g(a,b)|$ is obtained when $b = a + d$ with d such that $a = b + d[p]$ is also true. In this case, we have

$$|g(a,b)| \leq 2\left(2\left(1 - \frac{1}{p}\right)^2 + \frac{p-2}{p^2}\right) \leq 6,$$

and one can take $A_\mathbb{U} = 6$. Moreover, for all $a \in \{0, \dots, p-1\}$,

$$\begin{aligned} \mathbb{E}(g^2(U_i, a)) \leq& 2\mathbb{E}\left[\left(\sum_u \left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\left(\mathbf{1}_{a=u+d} - \frac{1}{p}\right)\right)^2\right] \\ &+ 2\mathbb{E}\left[\left(\sum_u \left(\mathbf{1}_{a=u} - \frac{1}{p}\right)\left(\mathbf{1}_{U_i=u+d} - \frac{1}{p}\right)\right)^2\right]. \end{aligned}$$

But

$$\mathbb{E}\left[\left(\sum_u \left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\left(\mathbf{1}_{a=u+d} - \frac{1}{p}\right)\right)^2\right]$$

$$= \mathbb{E}\left[\left(\left(\mathbf{1}_{U_i=a-d[p]} - \frac{1}{p}\right)\left(1 - \frac{1}{p}\right) - \frac{1}{p}\sum_{u\neq a-d[p]}\left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\right)^2\right].$$

Moreover the probability that $U_i = a - d[p]$ is $1/p$. Therefore, by straightforward computations,

$$\mathbb{E}\left[\left(\sum_u \left(\mathbf{1}_{U_i=u} - \frac{1}{p}\right)\left(\mathbf{1}_{a=u+d} - \frac{1}{p}\right)\right)^2\right]$$

$$= \frac{1}{p}\left(\left(1-\frac{1}{p}\right)^2 + \frac{p-1}{p^2}\right)^2 + \left(1-\frac{1}{p}\right)\left(-\frac{2}{p}\left(1-\frac{1}{p}\right) + \frac{p-2}{p^2}\right)^2$$

$$= \frac{1}{p}\left(1-\frac{1}{p}\right)^2 + \frac{1}{p^2}\left(1-\frac{1}{p}\right) = \frac{1}{p}\left(1-\frac{1}{p}\right) \leq \frac{1}{p}.$$

Therefore,

$$\mathbb{E}(g^2(U_i, a)) \leq \frac{4}{p}.$$

Hence, one can choose

$$C_{\mathbb{U}}^2 = \frac{2m(m-1)}{p} \quad \text{and} \quad B_{\mathbb{U}}^2 = \frac{4m}{p}.$$

Finally $D_{\mathbb{U}}$ is an upper bound over all functions a_i, b_j such that

$$\sum_{i=1}^{m-1} \mathbb{E}(a_i(U_i)^2) \leq 1 \quad \text{and} \quad \sum_{j=2}^{m} \mathbb{E}(b_j(U_j)^2) \leq 1$$

of

$$\begin{aligned}
\mathbb{E}\left[\sum_{i<j} a_i(U_i) g(U_i,U_j) b_j(U_j)\right] =& \mathbb{E}\left[\sum_{i=1}^{m-1} a_i(U_i) \sum_{j=i+1}^{m} \mathbb{E}(g(U_i,U_j)b_j(U_j)|U_j)\right] \\
\leq& \mathbb{E}\left[\sum_{i=1}^{m-1} |a_i(U_i)| \sum_{j=i+1}^{m} \sqrt{\mathbb{E}(b_j(U_j)^2)}\sqrt{\mathbb{E}(g(U_i,U_j)^2|U_j)}\right] \\
\leq& \frac{2}{\sqrt{p}}\mathbb{E}\left[\sum_{i=1}^{m-1} |a_i(U_i)| \sum_{j=i+1}^{m} \sqrt{\mathbb{E}(b_j(U_j)^2)}\right] \\
\leq& \frac{2\sqrt{m}}{\sqrt{p}}\mathbb{E}\left[\sum_{i=1}^{m-1} |a_i(U_i)|\right] \\
\leq& \frac{2m}{\sqrt{p}},
\end{aligned}$$

and $D_{\mathbb{U}} = \frac{2m}{\sqrt{p}}$ works. Therefore, by Theorem 3.4 of [55], for all $\theta > 0$,

$$\mathbb{P}(\mathbb{U}(d) \geq c(C_{\mathbb{U}}\sqrt{\theta} + D_{\mathbb{U}}\theta + B_{\mathbb{U}}\theta^{3/2} + A_{\mathbb{U}}\theta^2)) \leq 2.77e^{-\theta},$$

for c an absolute positive constant given in [55, 58]. A union bound gives the second result.

□

D.3 Proofs for data-dependent weights

Note that

$$\begin{aligned}\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*) =& \widetilde{A}^\top \left[\frac{I_p}{\sqrt{m}} - \frac{\sqrt{m}-1}{pm}\mathbb{1}_p\mathbb{1}_p^\top\right](Y - Ax^*) \\ =& \left[\frac{A^\top}{m} - \frac{m-1}{pm}\mathbb{1}_p\mathbb{1}_p^\top\right](Y - Ax^*).\end{aligned}$$

Therefore when applying the methodology of Section 3, we identify the ℓth component of R_k as

$$(R_k)_\ell = \frac{\mathbb{N}(\ell - k)}{m} - \frac{m-1}{pm}.$$

Thanks to this identification one can prove the following results.

Proposition D.2. *The constant weight given by (5.4) satisfy Assumption Weights(d) with probability larger than $1 - C/p$ for some absolute positive constant C.*

Proposition D.3. *Under the notations of Proposition D.2, there exists positive absolute constants c and C and an event of probability larger than $1 - C/p$ such that on this event*

$$d^2 \leq c\left(\frac{\log(p)^2}{p} + \frac{\log(p)^3}{m}\right)\left(\|x^*\|_1 + \frac{\log(p)}{m}\right).$$

Proposition D.4. *The non-constant weights given by (5.7) satisfy Assumption Weights(d) with probability larger than $1 - C/p$ for some absolute positive constant C.*

Proposition D.5. *Under the notations of Proposition D.2, there exists some absolute constants $\kappa_1, \kappa_2, c_1, c_2$ and C positive such that if $p \geq 5$ and if*

$$\kappa_1 \log(p)\sqrt{p} \leq m \leq \kappa_2 p \log(p)^{-1}, \tag{D.9}$$

there exists an event of probability larger than $1 - C/p$ such that on this event

$$c_1\left(\frac{x_k^* \log p}{m} + \frac{\log p}{p}\sum_{u \neq k} x_u^* + \frac{\log^2 p}{m^2}\right) \leq d_k^2 \leq c_2\left(\frac{x_k^* \log p}{m} + \frac{\log^2 p}{p}\sum_{u \neq k} x_u^* + \frac{\log^4 p}{m^2}\right).$$

D.3.1 Assumption Weights holds (proof of Propositions D.2 and D.4)

Proposition D.4 is just the application of (3.5) of Lemma 3.1 to each of the vectors R_k with $\theta = 2\log(p)$.

For Proposition D.2 note that

$$v_k = (R_k)_2^\top A x^* = \sum_{u=0}^{p-1} w(k-u) x_u^*.$$

Hence all the v_k's satisfy that

$$v_k \leq W \|x^*\|_1. \tag{D.10}$$

One can apply Lemma 3.1 with $R = -\mathbb{1}$ to obtain that

$$\mathbb{P}\left(-\overline{Y} \geq -\|x^*\|_1 + \sqrt{\frac{2}{m}\|x^*\|_1 \theta} + \frac{\theta}{3m} \right) \leq e^{-\theta},$$

which is equivalent to

$$\mathbb{P}\left(\|x^*\|_1 \geq \left[\sqrt{\frac{\theta}{2m}} + \sqrt{\frac{5\theta}{6m} + \overline{Y}} \right]^2 \right) \leq e^{-\theta}. \tag{D.11}$$

Therefore combining (3.3) of Lemma 3.1 with R_k with (D.10) and (D.11) leads to the desired result, taking $\theta = 2\log(p)$.

D.3.2 Bounds on the $\langle V_k, \mathbb{N} \rangle$s

Let $w(\ell) := \langle V_\ell, \mathbb{N} \rangle$. To derive bounds on the $w(\ell)$s, we need to introduce in addition to $\Omega_{\mathbb{U}}(\theta)$ another event, namely $\Omega_{\mathbb{N}}(\theta)$.

Lemma D.2. *There exists an event $\Omega_{\mathbb{N}}(\theta)$ of probability larger than $1 - 2pe^{-\theta}$ such that on $\Omega_{\mathbb{N}}(\theta)$, for all u in $\{0, \ldots, p-1\}$,*

$$\left| \mathbb{N}(u) - \frac{m}{p} \right| \leq \sqrt{2\frac{m}{p}\theta} + \frac{\theta}{3}.$$

This is just a classical consequence of Bernstein's inequality to the m i.i.d. variables $\mathbf{1}_{U_i = u}$. Thanks to this definition, one can prove the following bounds.

Lemma D.3. *There exists an absolute constant c such that for all $\theta > 1$, on the event $\Omega_{\mathbb{N}}(\theta)$, of probability larger than $1 - pe^{-\theta}$,*

$$W \leq c\left(\frac{\theta}{p} + \frac{\theta^2}{m} \right).$$

Proof of Lemma D.3. Recall that $W = \max w(\ell)$ with for fixed ℓ

$$w(\ell) = \sum_{u=0}^{p-1} \frac{1}{m^2} \left(\mathbb{N}(u) - \frac{m-1}{p} \right)^2 \mathbb{N}(u+\ell) = \langle V_\ell, \mathbb{N} \rangle.$$

Hence on $\Omega_{\mathbb{N}}(\theta)$

$$w(\ell) \leq \frac{1}{m^2} \left(\sqrt{2\frac{m}{p}\theta} + \frac{\theta}{3} + \frac{1}{p} \right)^2 \sum_{u=0}^{p-1} \mathbb{N}(u+\ell) \leq \square \frac{\mathbf{1}}{m} \left(\frac{m\theta}{p} + \theta^2 + \frac{1}{p^2} \right)^2.$$

But $1/(p^2 m) \leq \min(\theta/p, \theta^2/m)$, which gives the result. □

This bound can be refined for a particular range of values for m.

Lemma D.4. *If $p \geq 2$ and m satisfies*

$$5\max(2\kappa, 1)\theta\sqrt{p} \leq m \leq p\theta^{-1}, \tag{D.12}$$

then there exists positive constants c_1, c_2, c_1' and c_2' such that if $\theta > 3$, on $\Omega_{\mathbb{N}}(\theta) \cap \Omega_{\mathbb{U}}(\theta)$,

$$c_1/m \leq w(0) \leq c_2/m,$$

and for $\ell \neq 0$,

$$c_1'/p \leq w(\ell) \leq c_2'\theta/p.$$

Proof of Lemma D.4. Let $M(\theta) = m/p + \sqrt{2\frac{m}{p}\theta} + \frac{\theta}{3}$ be the bound given by Lemma D.2. For the upper bounds, first remark that

$$w(0) = \frac{1}{m^2}\sum_u \mathbb{N}(u)^3 - 2\frac{m-1}{pm^2}\sum_u \mathbb{N}(u)^2 + \left(\frac{m-1}{pm}\right)^2 \sum_u \mathbb{N}(u).$$

But $\sum_u \mathbb{N}(u) = m$ and on $\Omega_{\mathbb{N}}(\theta)$,

$$\begin{aligned}\sum_u \mathbb{N}(u)^3 \leq & \sum_{u/\mathbb{N}(u)\leq 1} \mathbb{N}(u) + \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u)^3 \\ \leq & \sum_{u/\mathbb{N}(u)\leq 1} \mathbb{N}(u) + M(\theta) \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u)^2 \\ \leq & \sum_{u/\mathbb{N}(u)\leq 1} \mathbb{N}(u) + M(\theta) \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u)(\mathbb{N}(u)-1) + M(\theta) \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u) \\ \leq & \sum_u \mathbb{N}(u) + (M(\theta)-1) \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u) + M(\theta) \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u)(\mathbb{N}(u)-1) \\ \leq & m + (2M(\theta)-1) \sum_{u/\mathbb{N}(u)>1} \mathbb{N}(u)(\mathbb{N}(u)-1).\end{aligned}$$

One can also write $\sum_u \mathbb{N}(u)^2 = m + \sum_u \mathbb{N}(u)(\mathbb{N}(u)-1)$. Therefore

$$w(0) \leq \frac{1}{m}\left(1 - \frac{m-1}{p}\right)^2 + \frac{1}{m^2}\left(2M(\theta) - 1 - 2\frac{m-1}{p}\right)\sum_u \mathbb{N}(u)(\mathbb{N}(u)-1).$$

But

$$\sum_u \mathbb{N}(u)(\mathbb{N}(u)-1) = \mathbb{U}(0) + \frac{m(m-1)}{p}$$

(see (D.5)). Therefore by Proposition D.1 on $\Omega_{\mathbb{N}}(\theta) \cap \Omega_{\mathbb{U}}(\theta)$

$$w(0) \leq \frac{1}{m}\left[\left(1 - \frac{m-1}{p}\right)^2 + \left(2M(\theta) - 1 - 2\frac{m-1}{p}\right)\left(\xi(\theta) + \frac{m-1}{p}\right)\right]. \tag{D.13}$$

But under (D.12), one has that

$$\xi(\theta) \leq 2\kappa\frac{\theta}{\sqrt{p}}$$

and

$$M(\theta) \leq K\theta,$$

for K an absolute constant large enough. Moreover, under (D.12), we observe that

$$\frac{\theta}{\sqrt{p}} \leq \frac{m}{p} \leq 1/\theta \leq 1.$$

This gives

$$w(0) \leq \frac{1}{m} + \square\frac{\theta}{p} + \square\frac{\theta^2}{m\sqrt{p}} \leq \frac{1}{m} + \square\frac{\theta}{p},$$

which gives the result since (D.12) holds.

Similarly, by using (D.4), for $d \neq 0$, on $\Omega_{\mathbb{N}}(\theta) \cap \Omega_{\mathbb{U}}(\theta)$,

$$\begin{aligned} m^2 w(d) =& \sum_u \mathbb{N}(u)^2 \mathbb{N}(u+d) - 2\frac{m-1}{p}\sum_u \mathbb{N}(u)\mathbb{N}(u+d) + \left(\frac{m-1}{p}\right)^2 \sum_u \mathbb{N}(u) \\ \leq& \left(M(\theta) - 2\frac{m-1}{p}\right)\left(\mathbb{U}(d) + \frac{m(m-1)}{p}\right) + m\left(\frac{m-1}{p}\right)^2 \\ \leq& m\left(M(\theta) - 2\frac{m-1}{p}\right)\left(\xi(\theta) + \frac{(m-1)}{p}\right) + m\left(\frac{m-1}{p}\right)^2. \end{aligned}$$

The same simplifications lead to the upper bound for $w(d)$.

For the lower bounds, remark that by the right hand side of (D.12), $(m-1)p^{-1} < 1/2$. Therefore

$$(\mathbb{N}(u) - (m-1)p^{-1})^2 \geq (1 - (m-1)p^{-1})^2,$$

for all $\mathbb{N}(u) \geq 1$ and therefore

$$w(0) \geq \frac{(1-(m-1)p^{-1})^2}{m^2} \sum_{u/\mathbb{N}(u)\geq 1} \mathbb{N}(u) = \frac{(1-(m-1)p^{-1})^2}{m} \geq \frac{1}{4m}.$$

If (D.12) is true,

$$\begin{aligned} m/5 \geq& \max(2\kappa, 1)\theta\sqrt{p} && \text{(D.14)} \\ \geq& \kappa\theta\sqrt{p} + \kappa\theta p p^{-1/2} && \text{(D.15)} \\ \geq& \kappa\theta\sqrt{p} + \kappa\theta^2\frac{5p}{m} && \text{(D.16)} \\ \geq& \kappa(\theta\sqrt{p} + \theta^2 p m^{-1}) = p\xi(\theta). && \text{(D.17)} \end{aligned}$$

But, by using (D.4), on $\Omega_{\mathbb{U}}(\theta)$, since $(m-1)p^{-1} < 1/3$,

$$\begin{aligned}
m^2 w(d) \geq & \sum_u \mathbb{N}(u)^2 \mathbb{N}(u+d) - 2\frac{m-1}{p}\sum_u \mathbb{N}(u)\mathbb{N}(u+d) + m\left(\frac{m-1}{p}\right)^2 \\
\geq & \left(1 - 2\frac{m-1}{p}\right)\mathbb{U}(d) + \frac{m(m-1)}{p}\left(1 - \frac{m-1}{p}\right) \\
\geq & -\left|1 - 2\frac{m-1}{p}\right| m\xi(\theta) + \frac{m(m-1)}{p}\left(1 - \frac{m-1}{p}\right) \\
\geq & \square m\left(\frac{m}{4p} - \xi(\theta)\right) \\
\geq & \square \frac{m^2}{20p}.
\end{aligned}$$

□

D.3.3 Control of the constant weight (proof of Proposition D.3)

Let $\theta > 1$. First remark that (3.2) with $R = \mathbb{1}$ gives that with probability larger than $1 - e^{-\theta}$

$$\bar{Y} \leq \square\left[\|x^*\|_1 + \frac{\theta}{m}\right].$$

Moreover using Lemma D.2, on $\Omega_{\mathbb{N}}(\theta)$,

$$B \leq \square\left[\sqrt{\frac{\theta}{mp}} + \frac{\theta}{m}\right].$$

Combining this with Lemma D.3 and taking $\theta = 2\log(p)$ gives

$$\begin{aligned}
d^2 \leq & \square\left[W\theta\left(\|x^*\|_1 + \frac{\theta}{m}\right) + \theta^2\left(\frac{\theta}{mp} + \frac{\theta^2}{m^2}\right)\right] \\
\leq & \square\left[\left(\frac{\theta}{p} + \frac{\theta^2}{m}\right)\theta\left(\|x^*\|_1 + \frac{\theta}{m}\right) + \theta^2\left(\frac{\theta}{mp} + \frac{\theta^2}{m^2}\right)\right],
\end{aligned}$$

which implies the result.

D.3.4 Control of the non-constant weights (proof of Proposition D.5)

Let $\theta = 2\log(p)$ (since $p \geq 5$, this ensures that $\theta > 3$). Applying (3.5) of Lemma 3.1 to $(R_k)_2$ gives that with probability larger than $1 - pe^{-\theta}$,

$$\hat{v}_k = \langle V_k, Y\rangle \leq \square\left[v_k + B^2\theta\right].$$

But since

$$v_k = \sum_u w(k-u)x^*_u,$$

one can use Lemma D.4 (by choosing κ_1, κ_2 such that (D.12) holds) to show that

$$\begin{aligned} d_k^2 &\leq \Box \left[v_k \theta + B^2 \theta^2 \right] \\ &\leq \Box \left[\frac{x_k^* \theta}{m} + \sum_{u \neq k} x_u^* \frac{\theta^2}{p} + \frac{\theta^4}{m^2} \right], \end{aligned}$$

since $\theta^2/m \geq \theta/p$.

For the lower bound, the arguments are similar

$$\begin{aligned} d_k^2 &\geq \Box \left[v_k \theta + B^2 \theta^2 \right] \\ &\geq \Box \left[\frac{x_k^* \theta}{m} + \sum_{u \neq k} x_u^* \frac{\theta}{p} + B^2 \theta^2 \right], \end{aligned}$$

but since $(m-1)/p < 1/3$ and since there is at least one $\mathbb{N}(u) \geq 1$ for some u, then $B > 2/3m^{-1}$. Hence

$$d_k^2 \geq \Box \left[\frac{x_k^* \theta}{m} + \sum_{u \neq k} x_u^* \frac{\theta}{p} + \frac{\theta^2}{m^2} \right],$$

which gives the result.

D.4 Proof of Proposition 5.4

Proof of Proposition 5.4. The first part of the proof follows the proof of Proposition 4.4, yielding

$$\|\widehat{x}^{\mathrm{OLS}} - x^*\|_2^2 = \|\widetilde{G}_{S^*}^{-1} \widetilde{A}_{S^*}^H (\widetilde{Y} - \widetilde{A} x^*)\|_2^2.$$

By Proposition 5.1, the maximum eigenvalue of $\widetilde{G}_{S^*}^{-1}$ is bounded by $\frac{1}{1-\sqrt{s^* \xi}} \leq c'$ under (5.5), yielding the result. □