A data-dependent weighted LASSO under Poisson noise

Xin Jiang, Patricia Reynaud-Bouret, Vincent Rivoirard, Laure Sansonnet, Rebecca Willett

▶ **To cite this version:**

Xin Jiang, Patricia Reynaud-Bouret, Vincent Rivoirard, Laure Sansonnet, Rebecca Willett. A data-dependent weighted LASSO under Poisson noise. 2015. hal-01226837

HAL Id: hal-01226837
https://hal.science/hal-01226837

Preprint submitted on 10 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A data-dependent weighted LASSO under Poisson noise

Xin Jiang, Patricia Reynaud-Bouret, Vincent Rivoirard, Laure Sansonnet, Rebecca Willett

September 29, 2015

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 weighted LASSO framework for broad classes of 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.

1 Introduction

Poisson noise arises in a wide variety of applications and settings, ranging among PET, SPECT, and pediatric or spectral CT [46, 25, 41] in medical imaging, x-ray astronomy [4, 3, 42], genomics [39], network packet analysis [16, 27], crime rate analysis [31], and social media analysis [47]. In these and other settings, observations are characterized by discrete counts of events (e.g. photons 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 [23, 33, 36, 32, 25, 42, 37].

Unfortunately, the development of risk bounds for sparse Poisson inverse problems presents some significant technical challenges. Methods that rely on the negative Poisson log-likehood 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 [23, 33, 36] 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) [43] or Compressed Sensing (CS) [10, 13]; 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 [2, 7, 26, 43, 45, 5]. 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 $A = (a_{kl})_{k,l} \in \mathbb{R}_+^{n\times p}$ and conditionally on A, we observe

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

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 tranformation 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 should be chosen to ensure Assumptions RE, Weights, and/or G (presented in Section 2) are satisfied to achieve the general results of Section 2. We provide explicit examples in Sections 4 and 5.

1.2 Weighted LASSO estimator for Poisson inverse problems

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

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

$$\widehat{x}^{\text{LS}} := I_{S^*}(\widetilde{A}_{S^*})^{\#}\widetilde{Y} \tag{1.1}$$

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. However, we show that the below LASSO estimators are able to correctly identify the support with high probability under certain conditions.

(Classical) LASSO estimator:

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

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

Weighted LASSO estimator:

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

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

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

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

where D is a diagonal matrix with the k^{th} diagonal element equal to d_k. 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 reweighing the columns of $\widetilde{A}$ to account for the heteroscedastic noise.

A weighted LASSO estimator similar to (1.3) has been proposed and analyzed in past work, notably [44, 2], where the weights are considered fixed and arbitrary. The analysis in [44], however, does not extend to signal-dependent noise (like 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 [2] or Restricted Isometry Property [8]. Similarly, the analysis of a weighted LASSO estimator described in [2] cannot directly be adopted in settings with signal-dependent noise and data-dependent weights.

If x^* has support of size s, then, over an appropriate range of values of s, the oracle estimator in (1.1) satisfies this very tight bound

$$\|\widehat{x}^{\mathrm{LS}} - x^*\|_2^2 \le \frac{1}{(1-\delta_s)^2} \sum_{k\in S^*} (\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k^2, \tag{1.5}$$

where δ_s is a parameter associated with the restricted eigenvalue condition of the sensing matrix $\widetilde{A}$ (see Proposition 2). Our analysis reveals that if we choose weights $d_1, \ldots, d_p$ satisfying

$$|(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k| \le d_k \qquad \text{for } k = 1, \ldots, p, \tag{1.6}$$

then similar results hold for the LASSO and weighted LASSO estimates under conditions of Theorem 3:

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_2^2 \le \frac{C_\gamma}{(1-\delta_s)^2} \sum_{k\in S^*} d_k^2 \tag{1.7}$$

and

$$\|\widehat{x}^{\mathrm{LASSO}} - x^*\|_2^2 \le \frac{C_\gamma}{(1-\delta_s)^2} s d^2, \tag{1.8}$$

where C_γ only depends on γ. Hence, if we do 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 to recover the exact same rates as the oracle, 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.

In practice the weights can only depend on the observed data: we show on two examples (a Bernoulli sensing matrix and random convolution, see Sections 4 and 5) that we can compute such weights from the data (by using Poisson concentration inequalities) such that (1.6) holds with high probability *and* those weights are small enough to ensure risk bounds consistent with prior art, leading to weighted LASSO estimates 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 [1, 19, 22, 48]. In some sense, the weights play the same role as the thresholds in the estimation procedure proposed in [14, 24, 34, 35, 39]. 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 understand the role of the weights more deeply, let us look at a very basic toy example where A is a diagonal matrix with decreasing eigenvalues $\lambda_1, \dots, \lambda_p$. It is well known that many classical inverse problems [12] can be rephrased as this toy example. Informally and since Poisson noise is just a particular case of heteroscedasticity, one could rephrase the direct problem in

$$y_k = \lambda_k x_k^* + \epsilon_k,$$

with ϵ_k of zero mean and standard deviation σ_k.

Let us first consider the ramifications of setting $\widetilde{Y} = Y$, $\widetilde{A} = A$. The quantity $1 - \delta_s$ appearing in (1.5) is then the smallest eigenvalue of A, that is, λ_p. On the other hand, by (1.6), d_k should be an upper bound on $\lambda_k \epsilon_k$. From a very heuristic point of view, d_k should therefore be of the order of $\lambda_k \sigma_k$ and the bound (1.5) is actually of the order of $\sum_{k \in S^*} \frac{\lambda_k^2 \sigma_k^2}{\lambda_p^2}$. In particular even if the true support S^* is included in the high values of the λ_k's , we still pay the worse case scenario with the division by λ_p^2. On the other hand, the classical inverse problem choice $\widetilde{Y} = A^{-1}Y$, $\widetilde{A} = A^{-1}A = I_p$ gives that $1 - \delta_s = 1$ and that d_k should be of the order of σ_k / λ_k. Therefore the upper bound (1.5) is then of the order of $\sum_{k \in S^*} \frac{\sigma_k^2}{\lambda_k^2}$. Then, for the interesting case where the $(\lambda_k^2)_{k \in S^*}$'s are larger than λ_p, this is much better and we only pay for ill-posedness in the support of x^*. 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 $\widetilde{Y}$ is really important for the rates.

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

In the two examples (Bernoulli and Convolution) of Sections 4 and 5, we need to choose $\widetilde{A}$ so that the corresponding Gram matrix $\tilde{G} = \widetilde{A}^\top \widetilde{A}$ is as close as possible to the identity (therefore $(1 - \delta_s)$ in (1.5) will be as close as possible to 1) 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 therefore giving the best possible rates. This choice in particular will enable us to get rates consistent with the minimax rates derived in [23] 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 (1.6). We then describe a general framework for finding such weights using the observed data in Section 3. We then 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 [15], and (b) a ill-posed Poisson deconvolution problem arising in genetic motif analysis, building upon the formulation described in [39]. We conclude with simulation-based verification of our derived rates.

2 Main result: Theoretical performance bounds for the weighted LASSO

In this section, the bounds that are derived do not take into account the noise structure and could be used whatever the underlying noise. They rely mostly on the following two main assumptions that are proved to be matched with high probability in the next sections. The first is known as the *Restricted Eigenvalue Condition* (see [2]).

ASSUMPTION RE(s, c_0) There exists $0 \leq \delta_{s,c_0} < 1$ such that for all $J \subset \{1, \ldots, p\}$ with $|J| \leq s$ and all $x \in \mathbb{R}^p$ satisfying

$$\|x_{J^c}\|_1 \leq c_0 \|x_J\|_1,$$

we have

$$\frac{\|\tilde{A}x\|_2^2}{\|x_J\|_2^2} \geq 1 - \delta_{s,c_0}.$$

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 $k = 1, \ldots, p$,

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)|_k \leq d_k. \tag{2.1}$$

In the sequel, we use the following notation:

$$d_{\max} := \max_{k \in \{1,\ldots,p\}} d_k, \qquad d_{\min} := \min_{k \in \{1,\ldots,p\}} d_k, \quad \text{and} \quad \rho_{\gamma,d} \geq \frac{d_{\max}}{d_{\min}} \frac{\gamma + 2}{\gamma - 2}. \tag{2.2}$$

Note that $\rho_{\gamma,d}$ is just a given upper bound on the ratio because $d_{\max}$ and $d_{\min}$ may depend on the underlying signal x^* and we would like to use a bound $\rho_{\gamma,d}$ that does not depend on the underlying signal, especially to prove that Assumption RE$(s, 2\rho_{\gamma,d})$ is matched with high probability whatever x^*.

2.1 Recovery error bounds

Our first result is a recovery error bound that does not assume sparsity on the underlying signal.

Theorem 1. *Let $s > 0$ be an integer and let x_s^* denote the best s-sparse approximation to x^*. Let $S^* := \mathrm{supp}(x_s^*)$. Let $\gamma > 2$ and assume that Assumption Weights$(\{d_k\}_k)$ is satisfied for some positive weights d_k's. Let $\rho_{\gamma,d}$ be consequently defined by (2.2). Define the bias term*

$$B_s := \max\{\|\widetilde{A}(x^* - x_s^*)\|_2^2, \|x^* - x_s^*\|_1\};$$

note that $B_s = 0$ when x^* is s-sparse. Under Assumption RE$(2s, 2\rho_{\gamma,d})$ with parameter $\delta_{2s,2\rho_{\gamma,d}}$, the Weighted LASSO estimator $\widehat{x}^{\mathrm{WL}}$ satisfies

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_2 \leq \frac{2\sqrt{2}\gamma(1+2\rho_{\gamma,d})}{1-\delta_{2s,2\rho_{\gamma,d}}} \left(B_s + \sum_{k\in S^*} d_k^2\right)^{1/2} + \left(3 + \frac{3}{(\gamma-2)d_{\min}}\right) B_s \tag{2.3}$$

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_1 \leq \frac{2\sqrt{2}\gamma(1+2\rho_{\gamma,d})}{1-\delta_{2s,2\rho_{\gamma,d}}} \left(B_s + \sum_{k\in S^*} d_k^2\right)^{1/2} \sqrt{s} + \left(3 + \frac{3}{(\gamma-2)d_{\min}}\right) B_s. \tag{2.4}$$

Remark 1. *Under the conditions of Theorem 1, we have for the LASSO estimator $\widehat{x}^{\mathrm{LASSO}}$*

$$\|x^* - \widehat{x}^{\mathrm{LASSO}}\|_2 \leq \frac{2\sqrt{2}\gamma(1+2\rho_{\gamma,d})}{1-\delta_{2s,2\rho_{\gamma,d}}} (B_s + d^2 s)^{1/2} + \left(3 + \frac{3}{(\gamma-2)d}\right) B_s \tag{2.5}$$

$$\|x^* - \widehat{x}^{\mathrm{LASSO}}\|_1 \leq \frac{2\sqrt{2}\gamma(1+2\rho_{\gamma,d})}{1-\delta_{2s,2\rho_{\gamma,d}}} (B_s + d^2 s)^{1/2} \sqrt{s} + \left(3 + \frac{3}{(\gamma-2)d}\right) B_s. \tag{2.6}$$

Note that we can take $\rho_{\gamma,d} = \frac{\gamma+2}{\gamma-2}$.

2.2 Support recovery guarantees

To obtain support recovery guarantees, we need the following much stronger condition.

Assumption G(ξ) Let $\widetilde{G} := \widetilde{A}^\top \widetilde{A}$ be the Gram matrix associated with $\widetilde{A}$. There exists a constant $\xi > 0$ such that

$$\left| \left(\widetilde{G} - I_p\right)_{k,\ell} \right| \leq \xi$$

for all $k, \ell \in \{1, \ldots, p\}$.

Assumption G(ξ) is stronger than RE(s, c_0) for sufficiently small values of ξ by the following result.

Proposition 1. *Under Assumption G(ξ) for all x vector of $\mathbb{R}^p$, $c_0 \geq 0$ and $J \subset \{1, \ldots, p\}$ such that*

$$\|x_{J^c}\|_1 \leq c_0 \|x_J\|_1,$$

the following inequality holds:

$$\frac{\|\tilde{A}x\|_2^2}{\|x_J\|_2^2} \geq 1 - (1+2c_0)|J|\xi.$$

Hence Assumption RE(s, c_0) holds with constant $\delta_{s,c_0} = (1+2c_0)\xi s$ as soon as

$$s(1+2c_0) < \xi^{-1}.$$

In particular, if $c_0 = 0$, namely x is supported by J, then we obtain the lower bound of a classical RIP as soon as

$$s < \xi^{-1}. \tag{2.7}$$

Theorem 2. *Assume that x^* is s-sparse for s a positive integer. Let $\gamma > 2$ and Assumptions RE$(s, 0)$ (with parameter $\delta_{s,0}$), Weights$(\{d_k\}_k)$, and G(ξ) be satisfied. Further assume s satisfies*

$$\xi \frac{2\gamma}{1-\delta_{s,0}} \left(s \sum_{k\in S^*} d_k^2\right)^{1/2} < \left(\frac{\gamma}{2} - 1\right) \min_{k\notin S^*} d_k. \tag{2.8}$$

Let $S^* = \mathrm{supp}(x^*)$. Under these conditions,

$$\mathrm{supp}(\widehat{x}^{\mathrm{WL}}) \subseteq \mathrm{supp}(x^*).$$

Theorem 2 is proved in Appendix A.3. This support guarantee is the main ingredient to derive sharper bounds than the ones of Theorem 1 at the price of in general a much stronger condition, namely Assumption G(ξ).

Note that an obvious choice for $\delta_{s,0}$ thanks to Proposition 1 is $\delta_{s,0} = s\xi$ but it is possible to have much better constants (see Section 4).

Remark 2. *In the special case where $d := d_1 = d_2 = \ldots = d_p$, Theorem 2 implies that if s is small enough, namely (since $s(1-\delta_{s,0})^{-1}$ is an increasing function of s) if s satisfies*

$$s(1-\delta_{s,0})^{-1} < \frac{\gamma - 2}{4\gamma}\xi^{-1}, \tag{2.9}$$

then LASSO estimator satisfies

$$\mathrm{supp}(\widehat{x}^{\mathrm{LASSO}}) \subseteq \mathrm{supp}(x^*).$$

2.3 Recovery error bounds for sparse signals

The recovery error bounds in Theorem 1 holds even for non-sparse x^*. However, if x^* is sufficiently sparse, even tighter bounds are possible, as we show below thanks to the support recovery guarantees. Before giving a better bound for the recovery error bounds than the one given in Theorem 1, let us look more closely at the oracle case.

Proposition 2. *Assume that $S^* = \mathrm{supp}(x^*)$ is known and $s = |S^*|$. If Assumptions RE(s, 0) (with parameter $\delta_{s,0}$) is satisfied; then $\widehat{x}^{\mathrm{LS}}$, the least square estimate of x^* on S^* satisfies*

$$\|\widehat{x}^{\mathrm{LS}} - x^*\|_2^2 \leq \frac{1}{(1-\delta_{s,0})^2} \sum_{k \in S^*} (\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k^2.$$

The weighted LASSO estimate satisfies the following bound.

Theorem 3. *Let $\gamma > 2$. Assume that x^* is s-sparse for s a positive integer satisfying the conditions of Theorem 2 and let $S^* := \mathrm{supp}(x^*)$. Then*

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_1 \leq \frac{2\gamma s^{1/2}}{1-\delta_{s,0}} \left(\sum_{k \in S^*} d_k^2 \right)^{1/2}, \tag{2.10}$$

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_2 \leq \frac{2\gamma}{1-\delta_{s,0}} \left(\sum_{k \in S^*} d_k^2 \right)^{1/2}, \tag{2.11}$$

and

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_\infty \leq \gamma d_{\max}.$$

Theorem 3 is proved in Appendix A.5.

Remark 3. *Similar results to those in Theorem 3 can be derived for the (unweighted) LASSO estimator by letting $d = d_1 = \ldots = d_p$.*

Remark 4. *We deduce that under assumptions of Theorem 3, for any $1 < q < \infty$, the weighted lasso estimate $\widehat{x}^{\mathrm{WL}}$ satisfies*

$$\|\widehat{x}^{\mathrm{WL}} - x^*\|_q^q \leq \frac{K_{\gamma,q} s}{1-\delta_{s,0}} d_{\max}^q,$$

for some constant $K_{\gamma,q}$ only depending on γ and q.

Remark 5. *Let us compare the assumptions of Theorem 1 and 2 under the result of Proposition 1. If Assumption G(ξ) is fulfilled and if we can take $\delta_{s,c_0} = s(1+2c_0)\xi < 1$ with $c_0 = 2\rho_{\gamma,d}$, the constraint on the sparsity level s for Theorem 1 can be rewritten as*

$$s\xi\left(1+4\frac{\gamma+2}{\gamma-2}\frac{d_{max}}{d_{min}}\right) < 1.$$

But on the other hand (2.8) is a consequence of

$$\xi\frac{2\gamma}{1-\delta_{s,0}} s d_{max} < \left(\frac{\gamma}{2}-1\right) d_{min},$$

with $\delta_{s,0} = s\xi$ thanks to Proposition 1. Hence this can be rewritten as

$$s\xi\left(1+4\frac{\gamma}{\gamma-2}\frac{d_{max}}{d_{min}}\right) < 1.$$

Therefore if Assumption G(ξ) holds and if we use $\delta_{s,2\rho_{\gamma,d}} = s(1+4\rho_{\gamma,d})\xi$ and not sharper constants, then the constraint on the sparsity is less stringent for Theorem 2 than for Theorem 1. Note that the fact that $\sum_{k\in S^*} d_k^2$ appears in (2.8) instead of its upper bound $s d_{max}^2$ makes in fact (2.8) much less troublesome to satisfy that the constraint coming from Assumption RE(s, c_0) (see in particular Section 5).

Remark 6. *We can also easily combine results of Theorems 2 and 3 to get exact support recovery as soon as*

$$\min_{k\in S^*} |x_k^*| > \bar{\gamma} d_{max}.$$

Indeed, if there exists k_1 such that $x_{k_1}^* \neq 0$ and $\widehat{x}_{k_1} = 0$, then

$$|x_{k_1}^*| \leq \|\widehat{x}^{\mathrm{LASSO}} - x^*\|_\infty \leq \gamma d_{max}$$

and we obtain a contradiction.

Remark 7. *Note that the bound in (2.11) is within a constant factor of the "optimal" rate associated with knowledge of the support of x^* in Proposition 2, as soon as the d_k's are chosen sharply enough to get $|(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*))_k| \simeq d_k$ whereas we lose at least a factor $(1+\rho_{\gamma,d})$ by only using Theorem 1, even if we assume that $\delta_{2s,2\rho_{\gamma,d}} \simeq \delta_{s,0}$. This is just a factor depending on γ if the weights are constant, but this can be much worse if the weights are non-constant and if $d_{\max} >> d_{\min}$.*

3 Choosing data-dependent weights

In general, choosing d_ks 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 chosen so that for all k

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

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 $R_k^\top$ 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. With the previous 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 1. *For all vectors $R = (R_\ell)_{\ell=1,\dots,n} \in \mathbb{R}^n$, eventually depending on A, let $R_2 = (R_\ell^2)_{\ell=1,\dots,n}$. 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

$$v = R_2^\top \mathbb{E}(Y|A) = R_2^\top Ax^*$$

and

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

Moreover

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

$$\mathbb{P}\left(v \geq \left(\sqrt{\frac{b^2\theta}{2}} + \sqrt{\frac{5b^2\theta}{6} + R_2^\top Y}\right)^2 \,\middle|\, A\right) \leq e^{-\theta} \tag{3.4}$$

and

$$\mathbb{P}\left(|R^\top Y - R^\top Ax*| \geq \left(\sqrt{\frac{b^2\theta}{2}} + \sqrt{\frac{5b^2\theta}{6} + R_2^\top Y}\right)\sqrt{2\theta} + \frac{b\theta}{3} \,\middle|\, A\right) \leq 3e^{-\theta}. \tag{3.5}$$

Equations (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 Equation (3.4) provides an estimated upper-bound for v with high probability. Equation (3.5) is therefore our main ingredient for giving observable d_k's that are satisfying Assumption Weights($\{d_k\}_k$). Note that depending on A, one may also find more particular way to define those weights, in particular constant ones. This is illustrated in the two following examples.

4 Example: 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 authors consider $q = 1/2$.) In fact, the celebrated Rice single-pixel camera [15] 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.* [23, 33]). In this section, we consider our proposed weighted-LASSO estimator for this sensing matrix.

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

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

satisfies $\mathbb{E}(\tilde{G}) = \mathbb{E}(\widetilde{A}^\top\widetilde{A}) = I_p$ (see Appendix), which will help us to ensure that Assumptions RE and G hold. To make d_k as small as possible and 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. Some computations given in the appendix shows that it is sufficient to take

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

In the below, we use the $\lesssim$, $\gtrsim$, and $\simeq$ notation to mask absolute constant factors.

4.2 Assumption RE holds with high probability

With high probability, the sensing matrix $\widetilde{A}$ considered here satisfies the Restricted Isometry property with parameter $\delta_s \leq 1/2$. That is:

Theorem 4 (RIP for Bernoulli sensing matrix (Theorem 2.4 in [30]))**.** *There exists constants $c_1, c_2, c_3, C > 0$ such that for any $\delta_s \in (0, 1/2)$, if*

$$s \leq \frac{c_1\delta_s^2 n}{\alpha_q^4 \log(c_2 p\alpha_q^4/\delta_s^2 n)}, \qquad \text{and} \qquad n \geq \frac{\alpha_q^4}{c_3\delta_s^2}\log p \tag{4.3}$$

where

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

then for any x -s-sparse,

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

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

Lemma 4.1(i) from [2], combined with Lemma 2.1 of [8], shows that Assumption RE(s, c_0) holds for δ_{s,c_0} with

$$\sqrt{1-\delta_{s,c_0}} = \sqrt{1-\delta_{2s}}\left(1 - \frac{\delta_{3s}c_0}{1-\delta_{2s}}\right).$$

Using $\delta_{3s} \geq \delta_{2s}$, straightforward computations show that Assumption RE(s, c_0) holds for δ_{s,c_0} as soon as $\delta_{s,c_0} \leq \delta_{3s}(1+2c_0)$. We use this result with $c_0 = 2\rho_{\gamma,d}$ and fix $0 < \delta_{6s} < (2+8\rho_{\gamma,d})^{-1}$. Then, as required by Theorem 1, Assumption RE($2s, 2\rho_{\gamma,d}$) holds for parameter $\delta_{2s,2\rho_{\gamma,d}} = \delta_{6s}(1+4\rho_{\gamma,d}) < 1/2$ with probability exceeding $1 - C/p$ for a universal positive constant C if

$$s \lesssim \frac{nq^2(1-q)^2}{\rho_{\gamma,d}^2 \log(p\rho_{\gamma,d}^2/nq^2(1-q)^2)}$$

and $n \gtrsim \frac{\rho_{\gamma,d}^2 \log p}{q^2(1-q)^2}$.

4.3 Assumption G holds with high probability

Proposition 3. *There is a universal positive constant C so that, with probability larger than $1 - C/p$,*

$$|(\widetilde{G} - I_p)_{k\ell}| \leq \xi,$$

with

$$\xi = \sqrt{\frac{6\log p}{n}\left(\frac{(1-q)^2}{q} + \frac{q^2}{1-q}\right)} + \frac{\log p}{n}\max\left(\frac{1-q}{q}, \frac{q}{1-q}\right).$$

4.4 Choice of the weights

Because the matrix A has a very special form, one can derive two sets of weights. Constant weights are easier to handle with respect to the previous Theorems because $d_{max}/d_{min} = 1$. However we will see that they may lead to some loss in the rates of convergence depending on the parameter values. Non-constant weights follow the general guidelines given by Lemma 1 and are therefore more easy to derive. However we need in this case to derive very precise upper and lower bounds to control d_{max}/d_{min}.

4.4.1 Constant weights

We can leverage the machinery described in Section 3 to derive an expression for a data-dependent weight d to use in the unweighted LASSO estimator defined in (1.2). Let

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

with

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

and let

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

be an estimator of $\|x^*\|_1$. Then the constant weights are given by

$$d = \sqrt{6W\log(p)}\sqrt{\hat{N}} + \frac{\log(p)}{(n-1)q(1-q)} + c\left(\frac{3\log(p)}{n} + \frac{9\max(q^2,(1-q)^2)}{n^2q(1-q)}\log(p)^2\right)\hat{N}, \tag{4.4}$$

where c is a constant and one can prove that they satisfy Assumption Weights(d) with high probability (see Proposition 5). Note that the third term in the expression for d in (4.4) is negligible when $\|x^*\|_1 \log p/n \ll 1$. Furthermore, as shown in Proposition 6, in the range

$$nq^2(1-q) \gtrsim \log(p)$$

we have

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

4.4.2 Non-constant weights

For all $k = 0, \dots, p-1$, define the vector V_k of size n as

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

The non-constant weights are given by

$$d_k = \sqrt{6\log(p)} \left(\sqrt{\frac{3\log(p)}{2(n-1)^2 q^2 (1-q)^2}} + \sqrt{\frac{5\log(p)}{2(n-1)^2 q^2 (1-q)^2} + V_k^\top Y} \right) + \frac{\log(p)}{(n-1)q(1-q)} + \tag{4.5}$$

$$+ c \left(\frac{3\log(p)}{n} + \frac{9\max(q^2, (1-q)^2)}{n^2 q(1-q)} \log(p)^2 \right) \hat{N}, \tag{4.6}$$

where c is a constant. They also satisfy Assumption Weights(d) with high probability (see Proposition 7). Furthermore, as shown in Proposition 8, in the range

$$nq^2(1-q) \gtrsim \log(p)$$

we have

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

4.5 Summary of rate results

In this section, we consider the data-dependent LASSO parameter d defined in (4.4) and the data-dependent weighted LASSO parameters $\{d_k\}_k$ defined in (4.6). In this setting, with probability exceeding $1 - C/p$ for a universal constant C

$$\frac{d_{\max}^2}{d_{\min}^2} \lesssim 1 + \frac{\|x^*\|_\infty}{\|x^*\|_1} \max(q^{-1}, (1-q)^{-1}) \lesssim \max(q^{-1}, (1-q)^{-1}) \tag{4.7}$$

so we may set, by using (2.2),

$$\rho_{\gamma,d} \simeq \frac{\gamma+2}{\gamma-2} \sqrt{\max(q^{-1}, (1-q)^{-1})}.$$

Fix $0 < \delta_{6s} < (2 + 8\rho_{\gamma,d})^{-1}$. If x^* is s-sparse, so that

$$s = |\operatorname{supp}(x^*)| \lesssim \frac{nq^2(1-q)^2}{\rho_{\gamma,d}^2 \log(p\rho_{\gamma,d}^2 / nq^2(1-q)^2)},$$

then we have the following with probability exceeding $1 - C/p$ for a universal constant C (neglecting constants depending on γ and $\delta_{s,0}$):

$$sd^2 \lesssim \frac{\log p}{n} \left(\frac{\|x^*\|_1 s}{\min(q, 1-q)} + \frac{\|x^*\|_1^2 s \log p}{n} + \frac{s \log p}{nq^2 \min(q, 1-q)^2} \right) \tag{4.8a}$$

$$\sum_{k \in \operatorname{supp}(x^*)} d_k^2 \lesssim \frac{\log p}{n} \left(\frac{\|x^*\|_1}{q} + \frac{\|x^*\|_1 s}{(1-q)} + \frac{\|x^*\|_1^2 s \log p}{n} + \frac{s \log p}{nq^2(1-q)^2} \right). \tag{4.8b}$$

We combine the above bounds values with Theorem 1 and Remark 1 to derive the following for $0 < q \leq 1/2$. (The case where $1/2 < q < 1$ can be examined using similar methods.) For

$$s \lesssim \frac{nq^3(1-q)^2}{\log(p/nq^3(1-q)^2)}$$

and

$$n \gtrsim \frac{\rho^2_{\gamma,d} \log p}{q^2(1-q)^2}$$

the oracle least-squares estimator on the true support of x^* satisfies

$$\|x^* - \widehat{x}^{\mathrm{LS}}\|_2^2 \lesssim \frac{\log p}{n}\left(\frac{\|x^*\|_1}{q} + \frac{\|x^*\|_1 s}{(1-q)} + \frac{\|x^*\|_1^2 s \log p}{n} + \frac{s \log p}{nq^2(1-q)^2}\right).$$

If we use the constant weight d, then for the same range of s and n we have

$$\|x^* - \widehat{x}^{\mathrm{LASSO}}\|_2^2 \lesssim \frac{\log p}{nq}\left(\frac{\|x^*\|_1 s}{q} + \frac{\|x^*\|_1^2 s \log p}{n} + \frac{s \log p}{nq^4}\right).$$

If we use the non-constant weights, then for the same range of s and n we have

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_2^2 \lesssim \frac{\log p}{nq}\left(\frac{\|x^*\|_1}{q} + \frac{\|x^*\|_1 s}{(1-q)} + \frac{\|x^*\|_1^2 s \log p}{n} + \frac{s \log p}{nq^2(1-q)^2}\right).$$

Note that the rates for $\widehat{x}^{\mathrm{LS}}$ and $\widehat{x}^{\mathrm{WL}}$ are equivalent up to the factor q^{-1} and lower than the rate for $\widehat{x}^{\mathrm{LASSO}}$. This suggests that when $q < 1/2$, for an appropriate range of s there can be some advantage to using the weighted LASSO in (1.3) over using the typical unweighted LASSO in (1.2). Furthermore, we note that the rates above are similar to the rates derived in a similar setting for estimators based on minimizing a regularized negative log-likelihood. Since the model in [23] considered x^* sparse in a basis different from the canonical or sampling basis and only considers $q = 1/2$, the two results are not perfectly comparable. Specifically, in [23], the sensing matrix was scaled differently to model certain physical constraints[1]; if we adjust the rates in [23] to account for this different scaling, we find a MSE decay rate of $s\|x^*\|_1 \log p/nq$ for $\|x^*\|_1$ sufficiently large, and that for smaller $\|x^*\|_1$ is MSE is constant with respect to $\|x^*\|_1$ but varies instead with p and n; those rates were validated experimentally in [23]. This shows the similarity between the rates derived with the proposed framework and previous results.

5 Example: Poisson random convolution in genomics

In this section we describe a specific random convolution model that is a toy model for bivariate Hawkes models or even more precise Poissonian interaction functions [19, 39, 40]. 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 [18, 11]. 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 in this case the corresponding Gram matrix G can be proved to be invertible. However, and

[1]Specifically, the sensing matrix considered in [23] corresponds to our A/qn and the data acquisition time T in [23] corresponds to our $nq\|x^*\|_1$. Thus the rates in [23] were essentially derived for the scaled MSE $n^2q^2\|x^* - \widehat{x}\|_2^2/T^2$.

in particular in genomics, it is not at all clear that such an horizon exists. Indeed it usually is assumed that the interaction stops after 10000 bases because the 3D structure of DNA makes the "linear distances" on the DNA not real anymore, but if one would have access to real 3D positions (and there are some data going in this direction), would it be possible to estimate the interaction functions without any assumption on its support? Of course, fully taking into account the 3D structure of the DNA might lead to other kind of difficulties, this is why we want to focus as a preliminary study, on the circle case.

Our toy model can be written in the following form. Let $U_1, \dots, U_m$ be a collection of m i.i.d. realizations of a Uniform random variable on the set $\{0, \dots, p-1\}$; these have to be thought of as points equally distributed on a circle. Each U_i will give birth independently to some Poisson variables. If $x^* = (x^*_0, \dots, x^*_{p-1})^\top$ is a vector in $\mathbb{R}^p_+$, then let us define $N^i_{U_i+j}$ to be a Poisson variable with law $\mathcal{P}(x^*_j)$ independent of anything else. The variable $N^i_{U_i+j}$ represents the number of children that a certain individual (parent) U_i has at distance j. 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. $Y_k = \sum_{i=1}^m N^i_k$. So the problem can be translated as follows: we observe the U_i's and the Y_k's whose law conditioned on the U_i's is given by

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

So this is a discretized version of the model given in [39] where the data have been binned and in addition put on a circle (which is also quite consistent with the fact that some bacteria genomes are circular). By adapting the assumptions of [39, 40, 19] to this set-up, the "classical" setting amounts to assume that the coordinates of x^* are null after a certain horizon s. So the main question is whether it is possible to estimate x^* if one only assumes that s coordinates are non zero but that we do not know where they are (i.e. not necessarily at the beginning of the sequence). The sparsity is a reasonable assumption in genomics for instance because if there is indeed a link between the parents and the children, then the main distances of interaction will correspond to particular chemical reactions going on.

The above model actually amounts to a random convolution, as detailed below. Other authors have studied random convolution, notably [38, 28, 20, 9], but those analyses do not extend to the problem considered here. For example, Candès and Plan [9, 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 line have coefficients with same magnitude. In our setting the Fourier coefficients of A are random and do not have uniform magnitude, and so the analysis in [9] 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

Put another way, we may write $Y := (Y_0, \dots, Y_{p-1})^\top$ and let $A \equiv A(U)$ be a $p \times p$ circulant matrix satisfying

$$A = \sum_{i=1}^m A_i, \qquad \text{where} \qquad A_i := \left[\begin{array}{cccccc} e_{U_i} & e_{U_i+1} & e_{U_i+2} & \cdots & e_{U_i+(p-1)} \end{array} \right], \tag{5.1}$$

where the subscripts are understood to be modulus p and e_i denotes the i^{th} column of a $p \times p$ identity matrix. We introduce the multinomial variable $\mathbb{N}$, defined for all $k \in \{0, \dots, p-1\}$ by

$$\mathbb{N}(k) = \text{card}\left\{i : \ U_i = k[p]\right\}. \tag{5.2}$$

It represents the number of parents at position k on the circle. Note that $\sum_{u=0}^{p-1} \mathbb{N}(u) = m$, which will be extensively used in proofs, and

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

that is, $A_{\ell,k} = \mathbb{N}(\ell - k)$. Using this notation, we are actually perfectly in the set-up of the present article, that is

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

Note that if it is a convolution model, it is actually a very particular one. Indeed, A is a square matrix satisfying $\mathbb{E}(A) = m\mathbb{1}_p\mathbb{1}_p^\top$ and therefore all the eigenvalues of the latter are null except the first one. In this sense it is a very badly ill-posed problem. This can also be viewed by the fact that in expectation, we are convoluting the data by a uniform distribution which is known to be a completely unsolvable problem. Therefore and as for many works on compressed sensing, we rely on the randomness to prove that Assumptions such as RE(s, c_0) are satisfied. Note that in some sense we are going further than [19] : in their case the Gram matrix was invertible because one knew where the non zero coefficients were, here Assumption RE(s, c_0) proves that we can somehow makes this property uniform whatever the (small) set of non zero coefficients is.

Finally, m the number of parents is somehow measuring the number observations, since we somehow get m copies of x^* plus some noise. Therefore a large m should improve rates. But if m tends to infinity, one should be close to the convolution by a uniform variable and we should not be able to recover the signal either, whereas if m is much smaller than p, there is "room" to see non overlapping shifted x^* and the estimation should be satisfying. This explains why our rates are clearly given by a competition between m and p as detailed hereafter.

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 $\mathbb{E}(\tilde{G}) = \mathbb{E}(\widetilde{A}^\top \widetilde{A}) = I_p$. Moreover for any $k \in \{0, \dots, p-1\}$, 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 Assumptions RE and G hold with high probability

The matrix $\tilde{G}$ can be reinterpreted thanks to U-statistics (see Proposition 9 in the appendix) and from this one can deduce the following result.

Proposition 4. *There exists absolute positive constants κ and C and an event of probability larger than $1 - C/p$, on which Assumption G(ξ) is satisfied with*

$$\xi := \kappa \left(\frac{\log p}{\sqrt{p}} + \frac{\log^2 p}{m} \right). \tag{5.3}$$

From Proposition 1 one can easily deduce that for all c_0, Assumption RE(s, c_0) is satisfied as long as

$$s < (1 + 2c_0)^{-1} \xi^{-1}$$

on an event of probability larger than $1 - C/p$.

Note that the rate of ξ is up to logarithmic terms $\max(p^{-1/2}, m^{-1})$ and therefore the sparsity level s can be large if both p and m tends to infinity. If we follow the intuition than m is the number of observations, and if only m tends to infinity, then we are rapidly limited by p meaning that even without the Poisson noise we would not be able to recover the whole signal x^*, if it has a very large support with respect to $\sqrt{p}$. We also see the reverse: if p grows but m is fixed then m is the limiting factor as usual for a fixed number of observations.

5.4 Choice of the weights

As for the Bernoulli case, one can have two choices: either constant weights that are using the very particular shape of A or non-constant weights that follows from Lemma 1.

5.4.1 Constant weights

To define the constant weights, let $W := \max_{\ell=0,\dots,p-1} w(\ell)$ with for all $\ell = 0, \dots, p-1$

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

and let

$$B = \max_{u \in \{0,\dots,p-1\}} \frac{1}{m} \left| \mathbb{N}(u) - \frac{m-1}{p} \right|.$$

Then the constant weights are given by

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

and one can prove that they satisfy Assumption Weights(d) with high probability (see Proposition 10). Furthermore, as shown in Proposition 11,

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

5.4.2 Non-constant weights

For all $k = 0, \dots, p-1$, the non-constant weights are given by

$$d_k = \sqrt{4 \log p} \left[\sqrt{\hat{v}_k + \frac{5B^2 \log p}{3}} + \sqrt{B^2 \log p} \right] + \frac{2B \log p}{3}, \tag{5.5}$$

with for all k in $\{0, \dots, p-1\}$,

$$\hat{v}_k = \sum_{\ell=1}^{p} \left(\mathbb{N}(\ell - k) - \frac{m-1}{p} \right)^2 \frac{Y_\ell}{m^2}.$$

They also satisfy Assumption Weights(d) with high probability (see Proposition 12). Furthermore, Proposition 13 shows that in the range

$$\log(p)\sqrt{p} \lesssim m \lesssim p \log(p)^{-1}. \tag{5.6}$$

we have

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

Note that the non-constant weights are interesting only if we are able to show that they significantly behave in a non-constant fashion. In Appendix D.3, we upper and lower bound them such that they match (2.8) with high probability in the range shown in (5.6) if

$$s \lesssim \frac{\sqrt{p}}{\log(p)^2}.$$

Note that this condition implies $s\xi \lesssim \log(p)^{-1}$.

5.5 Summary of rate results

For the convolution case, we have shown Assumption RE($s, 0$) holds as a consequence of Assumption G(ξ); thus, for s-sparse signals x^*, Theorem 3 gives better rates. As an illustration of the performance of the proposed approach for this deconvolution problem, we are able to derive the following bounds, that hold with high probability, as an easy consequence of Theorem 3 combined with the sharp bounds derived in Propositions 11 and 13 (neglecting constants depending on γ and $\delta_{s,0}$).

If we use the constant weight d, then as soon as

$$s \lesssim \min\left(\frac{\sqrt{p}}{\log(p)}, \frac{m}{\log(p)^2} \right),$$

we have

$$\|x^* - \widehat{x}^{\text{LASSO}}\|_2^2 \lesssim s \left(\frac{\log(p)^2}{p} + \frac{\log(p)^3}{m} \right) \left(\|x^*\|_1 + \frac{\log(p)}{m} \right).$$

In particular as soon as the signal strength $\|x^*\|_1$ is larger than $\log(p)/m$, this scales with $\|x^*\|_1$ and up to logarithmic term, the rate is of the order of $\min(p, m)^{-1} \times s\|x^*\|_1$.

If we use the non-constant weights, in the range (5.6) and under the following constraint on the sparsity level s,

$$s \lesssim \frac{\sqrt{p}}{\log(p)^2}$$

we can derive the following bounds, namely

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

In particular the weighted LASSO rate is clearly am improvement over the classical LASSO and proportional to $\frac{\log(p)\|x^*\|_1}{m}$ (i.e. independent of s) as soon as

$$s \lesssim \min\left(\frac{p}{m\log(p)}, \frac{m\|x^*\|_1}{\log(p)^3} \right).$$

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. More specifically, we simulate two-step estimators in which we first estimate the support using the (weighted) LASSO, and then compute a least-squares estimate on the estimated support. This approach, which has been analyzed in [44], reduces the bias of LASSO estimators. Specifically, we define

$$\widehat{S}^{\mathrm{LASSO}} = \left\{ i : |\widehat{x}^{\mathrm{LASSO}}| > 0 \right\} \tag{5.7a}$$

$$\widehat{x}^{\mathrm{LASSO-S}} = I_{\widehat{S}^{\mathrm{LASSO}}} (\widetilde{A}_{\widehat{S}^{\mathrm{LASSO}}})^{\#} \widetilde{y} \tag{5.7b}$$

$$\widehat{S}^{\mathrm{WL}} = \left\{ i : |\widehat{x}^{\mathrm{WL}}| > 0 \right\} \tag{5.7c}$$

$$\widehat{x}^{\mathrm{WL-S}} = I_{\widehat{S}^{\mathrm{WL}}} (\widetilde{A}_{\widehat{S}^{\mathrm{WL}}})^{\#} \widetilde{y} \tag{5.7d}$$

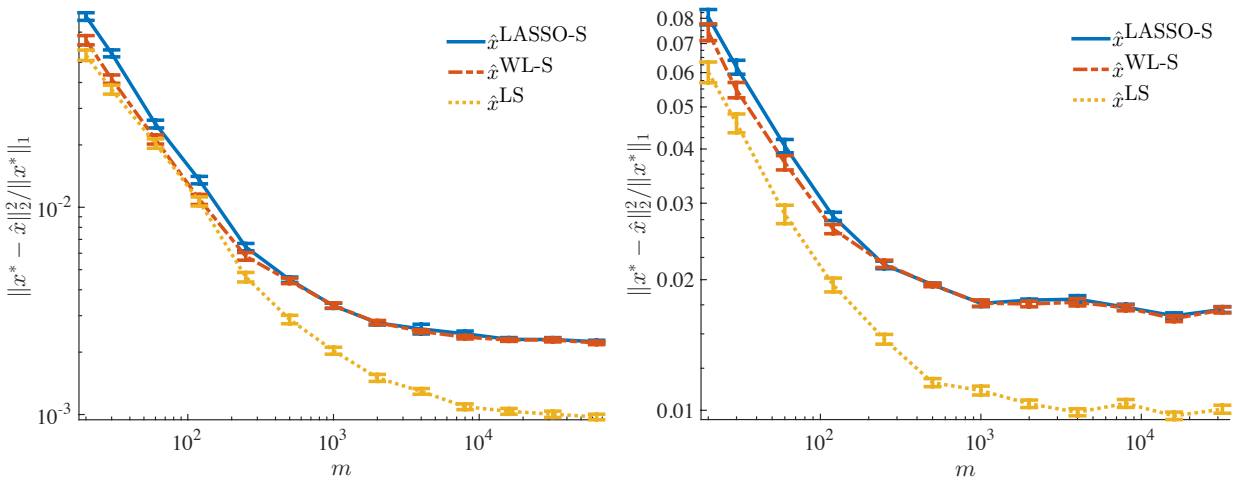
where the matrices $I_{\widehat{S}^{\mathrm{LASSO}}}$ and $I_{\widehat{S}^{\mathrm{WL}}}$ are used to fill in zeros at the off-support locations.

First we examine the MSE of $\widehat{x}^{\mathrm{LASSO-S}}$ and $\widehat{x}^{\mathrm{WL-S}}$ as a function of m, the number of uniform random events used to define the convolution operator A. We set $p = 5000$, and true support set of x^* is uniformly randomly selected for each experiment. The value of non-zeros are chosen from a exponential series (with a positive offset), and normalized so that the ℓ_1-norm of x^* are kept the same for all experiments. The tuning parameter γ is chosen to minimize the MSE. Each point in the plot is averaged over 400 random realizations. Figure 1 shows the normalized (over the ℓ_1-norm of signal) MSE as a functions of m for $s = 5$ and $s = 50$. Our rate shows that for both weighted LASSO and least-squares estimators, the MSE scales like $\frac{\|x^*\|_1 \log(p)}{m}$ when m is small and $\frac{s\|x^*\|_1 \log^2(p)}{p}$ when m is large relative to p; for the LASSO estimator, the MSE scales like $\frac{s\|x^*\|_1 \log^3(p)}{m}$ when m is small relative to p and $\frac{s\|x^*\|_1 \log^2(p)}{p}$ when m is large relative to p. Thus the weighted LASSO and least-squares estimators should outperform the LASSO estimator by a factor of s for small m.

Next we examine the MSE of $\widehat{x}^{\mathrm{LASSO-S}}$ and $\widehat{x}^{\mathrm{WL-S}}$ as a function of p, the length of x^*. In this experiment, we set $s = 5$, the true support set of x^* is uniformly randomly selected for each experiment. The value of non-zeros are chosen from a exponential series (with a positive offset). For each p, we set $m \propto \sqrt{p}\log(p)$ (thus m varies from 17 to 46 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. Note that when $m \propto \sqrt{p}\log(p)$ and $s \lesssim \sqrt{p}/\log^2 p$, the error rate for the weighted-LASSO is $\|x^*\|_1/\sqrt{p}$ while the error rate for LASSO is $\|x^*\|_1 s\log^2 p/\sqrt{p}$, which is worse by a factor of $s\log^2 p$.

6 Discussion and Conclusions

The data-dependent weighted LASSO method presented in this paper is a novel approach to sparse inference in the presence of heteroscedastic noise. We show that a simple assumption on the weights



(a) $s = 5$ (b) $s = 50$

Figure 1: MSE vs. m for the estimators in (1.1) and (5.7). The (oracle) least-squares estimator yields the smallest estimation error,as expected. When m is small, the weighted LASSO estimator outperforms the standard LASSO estimator, as predicted by the theory. When m is large, weighted LASSO and LASSO perform similarly. This is understandable because the variation of $\hat{d}_k$ becomes smaller as m grows, thus for very large values of m, the weighted LASSO and LASSO estimators are almost equivalent.

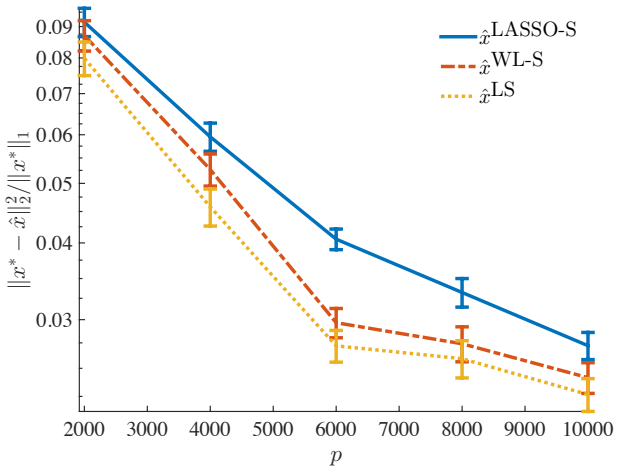


Figure 2: MSE vs. p with $m \propto \sqrt{p} \log p$ for the estimators in (1.1) and (5.7).

leads to estimation errors which are within a constant factor of the 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 [23], 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 [23].

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 alternative photon sources. 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 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.

It is worth noting that the core results of Section 2 do not use any probabilist arguments and therefore do not rely at all on Poisson noise assumptions. The Poisson noise model is only used to check that the necessary assumptions are fulfilled 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 and/or G.
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. *Therefore, the proposed weighted LASSO framework has potential for a variety of settings and noise distributions.*

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.

References

[1] Bertin, K., Le Pennec, E. & Rivoirard, V. (2011) Adaptive Dantzig density estimation. in *Annales de l'institut Henri Poincaré (B)*, vol. 47, pp. 43–74.

[2] Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, pp. 1705–1732.

[3] Borkowski, K. J., Reynolds, S. P., Green, D. A., Hwang, U., Petre, R., Krishnamurthy, K. & Willett, R. (2013) Supernova Ejecta in the Youngest Galactic Supernova Remnant G1.9+0.3. *The Astrophysical Journal Letters*, **771**(1), arXiv:1305.7399.

[4] ——— (2014) Nonuniform Expansion of the Youngest Galactic Supernova Remnant G1.9+0.3. *The Astrophysical Journal Letters*, **790**(2), arXiv:1406.2287.

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

[6] Bunea, F. (2008) Consistent selection via the Lasso for high dimensional approximating regression models. in *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, pp. 122–137. Institute of Mathematical Statistics.

[7] Bunea, F., Tsybakov, A. B. & Wegkamp, M. H. (2007) Aggregation for Gaussian regression. *Ann. Statist.*, **35**(4), 1674–1697.

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

[9] Candes, E. J. & Plan, Y. (2011) A probabilistic and RIPless theory of compressed sensing. *Information Theory, IEEE Transactions on*, **57**(11), 7235–7254.

[10] Candes, E. J. & Tao, T. (2006) Near-optimal signal recovery from random projections: Universal encoding strategies?. *Information Theory, IEEE Transactions on*, **52**(12), 5406–5425.

[11] Carstensen, L., Sandelin, A., Winther, O. & Hansen, N. R. (2010) Multivariate Hawkes process models of the occurrence of regulatory elements and an analysis of the pilot ENCODE regions. *BMC Bioinformatics*, **11**(456).

[12] Cavalier, L. (2011) Inverse problems in statistics. in *Inverse problems and high-dimensional estimation*, vol. 203 of *Lect. Notes Stat. Proc.*, pp. 3–96. Springer, Heidelberg.

[13] Donoho, D. L. (2006) Compressed Sensing. *IEEE Transactions on Information Theory*, **52**(4), 1289–1306.

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

[15] Duarte, M. F., Davenport, M. A., Takhar, D., Laska, J. N., Sun, T., Kelly, K. F. & Baraniuk, R. G. (2008) Single Pixel Imaging via Compressive Sampling. *IEEE Sig. Proc. Mag.*, **25**(2), 83–91.

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

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

[18] Gusto, G. & Schbath, S. (2005) 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).

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

[20] Harmany, Z. T., Marcia, R. F. & Willett, R. M. (2011) Spatio-temporal compressed sensing with coded apertures and keyed exposures. *arXiv preprint arXiv:1111.7247*.

[21] Houdré, C. & Reynaud-Bouret, P. (2003) Exponential inequalities, with constants, for U-statistics of order two. in *Stochastic inequalities and applications*, pp. 55–69. Springer.

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

[23] Jiang, X., Raskutti, G. & Willett, R. (2014) Minimax Optimal Rates for Poisson Inverse Problems with Physical Constraints. To appear in *IEEE Transactions on Information Theory*, arXiv:1403:6532.

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

[25] Lingenfelter, D. J., Fessler, J. A. & He, Z. (2009) Sparsity regularization for image reconstruction with Poisson data. in *IS&T/SPIE Electronic Imaging*, pp. 72460F–72460F. International Society for Optics and Photonics.

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

[27] Lu, Y., Montanari, A., Prabhakar, B., Dharmapurikar, S. & Kabbani, A. (2008) Counter Braids: A novel counter architecture for per-flow measurement. in *Proc. ACM SIGMETRICS*.

[28] Marcia, R. F. & Willett, R. M. (2008) Compressive coded aperture superresolution image reconstruction. in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pp. 833–836. IEEE.

[29] Massart, P. (2007) *Concentration inequalities and model selection*, vol. 6. Springer.

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

[31] Osgood, D. W. (2000) Poisson-Based Regression Analysis of Aggregate Crime Rates. *Journal of Quantitative Criminology*, **16**(1).

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

[33] Raginsky, M., Willett, R. M., Harmany, Z. T. & Marcia, R. F. (2010) Compressed sensing performance bounds under Poisson noise. *Signal Processing, IEEE Transactions on*, **58**(8), 3990–4002.

[34] Reynaud-Bouret, P. & Rivoirard, V. (2010) Near optimal thresholding estimation of a Poisson intensity on the real line. *Electronic journal of statistics*, **4**, 172–238.

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

[36] Rish, I. & Grabarnik, G. (2009) Sparse Signal Recovery with Exponential-Family Noise. in *Allerton Conference on Communication, Control, and Computing*.

[37] Rohban, M. H., Motamedvaziri, D. & Saligrama, V. (2015) Minimax Optimal Sparse Signal Recovery with Poisson Statistics. *arXiv preprint arXiv:1501.05200*.

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

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

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

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

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

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

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

[45] Van de Geer, S. A. (2008) High-dimensional generalized linear models and the LASSO. *The Annals of Statistics*, **36**(2), 614–645.

[46] Willett, R. M. & Nowak, R. D. (2003) Platelets: a multiscale approach for recovering edges and surfaces in photon-limited medical imaging. *Medical Imaging, IEEE Transactions on*, **22**(3), 332–350.

[47] Xu, J.-M., Bhargava, A., Nowak, R. & Zhu, X. (2012) Socioscope: Spatio-Temporal Signal Recovery from Social Media. *Machine Learning and Knowledge Discovery in Databases*, **7524**.

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

A Proofs of the LASSO bounds of Section 2

In the sequel, we shall use the following classical lemma.

Lemma A.1. *By the first-order optimality condition, we have that $\widehat{x}$ is a minimizer of $C(x) := \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=1}^p d_k |x_k|$ if and only if*

$$(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}))_k = \frac{\gamma d_k}{2} \operatorname{sign}(\widehat{x}_k) \qquad \text{for } k \text{ s.t. } \widehat{x}_k \neq 0 \tag{A.1a}$$

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x})|_k \leq \frac{\gamma d_k}{2} \qquad \text{for } k \text{ s.t. } \widehat{x}_k = 0. \tag{A.1b}$$

Before proceeding to recovery bounds, we establish the following oracle inequality.

Theorem A.1. *Let $\gamma > 2$, and $d_k > 0$, $k = 1, \dots, p$, such that Assumption Weights($\{d_k\}_k$) holds. Let $\rho = \rho_{\gamma,d}$ from (2.2). For any $s \in \{1, \dots, p\}$, let $\delta_{s,2\rho} > 0$ be a parameter such that Assumption RE($s, 2\rho$) holds. Then the weighted LASSO estimate $\widehat{x}^{\mathrm{WL}}$ defined in (1.3) satisfies*

$$\|\widetilde{A}\widehat{x}^{\mathrm{WL}} - \widetilde{A}x^*\|_2^2 \leq 8\gamma^2 \inf_{x:|\operatorname{supp}(x)|\leq s} \left\{ \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + \frac{\sum_{k\in\operatorname{supp}(x)} d_k^2}{1-\delta_{s,2\rho}} \right\}. \tag{A.2}$$

Furthermore, if x^* is s-sparse, under the same assumptions, (A.2) implies

$$\|\widetilde{A}\widehat{x}^{\mathrm{WL}} - \widetilde{A}x^*\|_2^2 \leq \frac{8\gamma^2}{1-\delta_{s,2\rho}} \sum_{k\in\operatorname{supp}(x^*)} d_k^2. \tag{A.3}$$

Theorem A.1 is proved in the sequel of Appendix A.

Remark A.1. *Under the assumptions of Theorem A.1, in the special case where $d := d_1 = d_2 = \ldots = d_p$, (A.2) implies that the LASSO estimate $\widehat{x}^{\mathrm{LASSO}}$ defined in (1.2) satisfies*

$$\|\widetilde{A}\widehat{x}^{\mathrm{LASSO}} - \widetilde{A}x^*\|_2^2 \leq 8\gamma^2 \inf_{x:|\operatorname{supp}(x)|\leq s} \left\{ \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + \frac{sd^2}{1-\delta_{s,2\rho}} \right\} \tag{A.4}$$

for any $s \in \{1, \dots, p\}$. Furthermore, if x^* is s-sparse, under the same assumptions,

$$\|\widetilde{A}\widehat{x}^{\mathrm{LASSO}} - \widetilde{A}x^*\|_2^2 \leq \frac{8\gamma^2}{1-\delta_{s,2\rho}} sd^2. \tag{A.5}$$

For short, in this section, we denote $\widehat{x}$ instead of $\widehat{x}^{\mathrm{WL}}$. Therefore,

$$\widehat{x} := \operatorname*{argmin}_{x\in\mathbb{R}^p} C(x).$$

To prove Theorem 2, we introduce the pseudo-estimate $\widehat{x}^{(S^*)}$ defined by

$$\widehat{x}^{(S^*)} \in \operatorname*{argmin}_{x\in\mathbb{R}^p:\operatorname{supp}(x)\subseteq S^*} \left\{ \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=1}^p d_k|x_k| \right\}. \tag{A.6}$$

We first state the following technical lemma.

Lemma A.2. *Let $x \in \mathbb{R}^p$. If Assumption Weights($\{d_k\}_k$) holds, we have*

$$(\gamma-2) \sum_{k\notin\operatorname{supp}(x)} d_k|\widehat{x}_k - x_k| \leq \|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 + (\gamma+2) \sum_{k\in\operatorname{supp}(x)} d_k|\widehat{x}_k - x_k|, \tag{A.7}$$

and if $\gamma > 2$, then

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma \sum_{k\in\operatorname{supp}(x)} d_k|\widehat{x}_k - x_k|. \tag{A.8}$$

Similarly for $\widehat{x}^{(S^*)}$, for any x supported by S^*, we have

$$(\gamma-2) \sum_{k\notin\operatorname{supp}(x)} d_k|\widehat{x}_k^{(S^*)} - x_k| \leq \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + (\gamma+2) \sum_{k\in\operatorname{supp}(x)} d_k|\widehat{x}_k^{(S^*)} - x_k|, \tag{A.9}$$

and if $\gamma > 2$, then

$$\|\widetilde{A}\widehat{x}^{(S^*)} - \widetilde{A}x^*\|_2^2 \leq \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma \sum_{k\in\operatorname{supp}(x)} d_k|\widehat{x}_k^{(S^*)} - x_k|. \tag{A.10}$$

Proof of Lemma A.2. By the optimality of $\widehat{x}$ according to (1.3), we have for any $x \in \mathbb{R}^p$,

$$\|\widetilde{Y} - \widetilde{A}\widehat{x}\|_2^2 + \gamma \sum_{k=1}^{p} d_k |\widehat{x}_k| \leq \|\widetilde{Y} - \widetilde{A}x\|_2^2 + \gamma \sum_{k=1}^{p} d_k |x_k|.$$

We obtain:

$$\begin{aligned}
&\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \qquad \text{(A.11)} \\
=&\|\widetilde{A}\widehat{x} - \widetilde{Y}\|_2^2 + \|\widetilde{Y} - \widetilde{A}x^*\|_2^2 + 2\langle \widetilde{A}\widehat{x} - \widetilde{Y}, \widetilde{Y} - \widetilde{A}x^* \rangle + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \\
\leq&\|\widetilde{Y} - \widetilde{A}x^*\|_2^2 + 2\langle \widetilde{A}\widehat{x} - \widetilde{Y}, \widetilde{Y} - \widetilde{A}x^* \rangle + \|\widetilde{Y} - \widetilde{A}x\|_2^2 \\
&+ \gamma \sum_{k=1}^{p} d_k |x_k| - \gamma \sum_{k=1}^{p} d_k |\widehat{x}_k| + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \\
=&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\langle \widetilde{A}\widehat{x} - \widetilde{A}x, \widetilde{Y} - \widetilde{A}x^* \rangle + \gamma \sum_{k=1}^{p} d_k (|x_k| - |\widehat{x}_k|) + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \\
=&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\langle \widehat{x} - x, \widetilde{A}^\top (\widetilde{Y} - \widetilde{A}x^*) \rangle + \gamma \sum_{k=1}^{p} d_k (|x_k| - |\widehat{x}_k|) + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \\
\leq&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2 \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| + \gamma \sum_{k=1}^{p} d_k (|x_k| - |\widehat{x}_k|) + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \\
\leq&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + \gamma \sum_{k=1}^{p} d_k (|x_k| - |\widehat{x}_k| + |x_k - \widehat{x}_k|) \\
\leq&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma \sum_{k \in \mathrm{supp}(x)} d_k |x_k - \widehat{x}_k|. \qquad \text{(A.12)}
\end{aligned}$$

Now, note that for any $x \in \mathbb{R}^p$,

$$\begin{aligned}
(\gamma - 2) \sum_{k \notin \mathrm{supp}(x)} d_k |\widehat{x}_k - x_k| \leq&\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| - (\gamma - 2) \sum_{k \in \mathrm{supp}(x)} d_k |\widehat{x}_k - x_k| \\
\leq&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma \sum_{k \in \mathrm{supp}(x)} d_k |\widehat{x}_k - x_k| - (\gamma - 2) \sum_{k \in \mathrm{supp}(x)} d_k |\widehat{x}_k - x_k| \\
\leq&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + (\gamma + 2) \sum_{k \in \mathrm{supp}(x)} d_k |\widehat{x}_k - x_k|,
\end{aligned}$$

and we obtain (A.7). For $\gamma \geq 2$ we have

$$\begin{aligned}
\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq&\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 + (\gamma - 2) \sum_{k=1}^{p} d_k |\widehat{x}_k - x_k| \\
\leq&\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma \sum_{k \in \mathrm{supp}(x)} d_k |\widehat{x}_k - x_k|
\end{aligned}$$

by (A.11) yielding (A.8). By using similar arguments, we obtain (A.9) and (A.10) by assuming that x is supported by S^*. □

Proof of Theorem A.1. Let $x \in \mathbb{R}^p$ such that $J = \mathrm{supp}(x)$ satisfies $|J| \leq s$. Let $\Delta = D(\widehat{x} - x)$, with D a diagonal matrix and for any k, $D_{kk} = d_k$. Our proof structure is inspired by the proof of Theorem 6.1 of [2]. By using (A.8), we have

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma\|\Delta_J\|_1. \tag{A.13}$$

If

$$2\gamma\|\Delta_J\|_1 \leq \frac{2\gamma}{\gamma+2}\|\widetilde{A}x - \widetilde{A}x^*\|_2^2,$$

then the proof follows easily. So we assume that

$$\frac{1}{\gamma+2}\|\widetilde{A}x - \widetilde{A}x^*\|_2^2 < \|\Delta_J\|_1.$$

By (A.13), we have

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma\left(\sum_{k\in J} d_k^2\right)^{1/2}\|(\widehat{x}-x)_J\|_2$$

and we have to bound $\|(\widehat{x}-x)_J\|_2$. Using (A.7),

$$\begin{aligned}(\gamma-2)\|\Delta_{J^c}\|_1 \leq& \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + (\gamma+2)\|\Delta_J\|_1 &\text{(A.14)}\\ \leq& (2\gamma+4)\,\|\Delta_J\|_1. &\text{(A.15)}\end{aligned}$$

We deduce that

$$\|(\widehat{x}-x)_{J^c}\|_1 \leq 2\rho\|(\widehat{x}-x)_J\|_1.$$

By Assumption RE, we obtain:

$$\|(\widehat{x}-x)_J\|_2^2 \leq \frac{\|\widetilde{A}(\widehat{x}-x)\|_2^2}{1-\delta_{s,2\rho}}. \tag{A.16}$$

Using the triangle inequality and the standard inequality $2ab \leq a^2/\tau + \tau b^2$ with $0 < \tau < 1$, $a, b > 0$, we obtain

$$\begin{aligned}\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq& \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma\left(\sum_{k\in J} d_k^2\right)^{1/2}\|(\hat{x}-x)_J\|_2 &\text{(A.17)}\\ \leq& \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma\left(\sum_{k\in J}\frac{d_k^2}{1-\delta_{s,2\rho}}\right)^{1/2}\|\widetilde{A}(\widehat{x}-x)\|_2 &\text{(A.18)}\\ \leq& \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + 2\gamma\left(\sum_{k\in J}\frac{d_k^2}{1-\delta_{s,2\rho}}\right)^{1/2}\left(\|\widetilde{A}(\widehat{x}-x^*)\|_2 + \|\widetilde{A}(x^*-x)\|_2\right) &\text{(A.19)}\\ \leq& \|\widetilde{A}x - \widetilde{A}x^*\|_2^2 + \gamma\left(\frac{2}{\tau(1-\delta_{s,2\rho})}\sum_{k\in J} d_k^2 + \tau\|\widetilde{A}(\widehat{x}-x^*)\|_2^2 + \tau\|\widetilde{A}(x^*-x)\|_2^2\right); &\text{(A.20)}\end{aligned}$$

letting $\tau = 1/(2\gamma)$ gives the result. □

A.1 Proof of Theorem 1

In this section we prove the following statement, from which Theorem 1 easily follows:

Theorem A.2. *Let $s > 0$ be an integer and let x_s^* denote the best s-sparse approximation to x^*. Let $S^* := \mathrm{supp}(x_s^*)$. Let $\gamma > 2$ and assume that Assumption Weights($\{d_k\}_k$) is satisfied for some positive weights d_k's. Let $\rho_{\gamma,d}$ be consequently defined by (2.2). Under Assumption RE($2s, 2\rho_{\gamma,d}$) with parameter $\delta_{2s,2\rho_{\gamma,d}}$, the Weighted LASSO estimator $\widehat{x}^{\mathrm{WL}}$ satisfies*

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_2 \leq \frac{2\sqrt{2}\gamma(1+2\rho_{\gamma,d})}{1-\delta_{2s,2\rho_{\gamma,d}}} \left(\|\widetilde{A}(x^* - x_s^*)\|_2^2 + \sum_{k\in S^*} d_k^2 \right)^{1/2} \tag{A.21}$$

$$+ 3\|x^* - x_s^*\|_1 + \left(\frac{2}{(\gamma-2)d_{\min}} + \frac{1}{(\gamma+2)d_{\max}} \right) \|\widetilde{A}(x^* - x_s^*)\|_2^2 \tag{A.22}$$

$$\|x^* - \widehat{x}^{\mathrm{WL}}\|_1 \leq \frac{2\sqrt{2}\gamma(1+2\rho_{\gamma,d})}{1-\delta_{2s,2\rho_{\gamma,d}}} \left(\|\widetilde{A}(x^* - x_s^*)\|_2^2 + \sum_{k\in S^*} d_k^2 \right)^{1/2} \sqrt{s} \tag{A.23}$$

$$+ 3\|x^* - x_s^*\|_1 + \left(\frac{2}{(\gamma-2)d_{\min}} + \frac{1}{(\gamma+2)d_{\max}} \right) \|\widetilde{A}(x^* - x_s^*)\|_2^2. \tag{A.24}$$

Proof of Theorem A.2. For simplicity of notation, we denote $\rho_{\gamma,d}$ as ρ in the following. Let T_0 be the indices of the s largest magnitude elements of x^* and T_1 be the indicies of the s largest magnitude elements of $h := x^* - \widehat{x}$ not in T_0. Let $T_{01} = T_0 \cup T_1$. Lemma A.2 with $x = x^*_{T_0}$ gives

$$(\gamma-2)d_{\min}\|h_{T_0^c}\|_1 \leq (\gamma-2)d_{\min}\|x^*_{T_0^c}\|_1 + \|\widetilde{A}x^*_{T_0^c}\|_2^2 + (\gamma+2)d_{\max}\|h_{T_0}\|_1.$$

If $(\gamma-2)d_{\min}\|x^*_{T_0^c}\|_1 + \|\widetilde{A}x^*_{T_0^c}\|_2^2 > (\gamma+2)d_{\max}\|h_{T_0}\|_1$, then $\|h_{T_0^c}\|_1 \leq 2\|x^*_{T_0^c}\|_1 + \frac{2}{(\gamma-2)d_{\min}}\|\widetilde{A}x^*_{T_0^c}\|_2^2$ and

$$\|h\|_2 \leq \|h\|_1 = \|h_{T_0^c}\|_1 + \|h_{T_0}\|_1 \leq \left(2 + \frac{(\gamma-2)d_{\min}}{(\gamma+2)d_{\max}}\right)\|x^*_{T_0^c}\|_1 + \left(\frac{2}{(\gamma-2)d_{\min}} + \frac{1}{(\gamma+2)d_{\max}}\right)\|\widetilde{A}x^*_{T_0^c}\|_2^2. \tag{A.25}$$

Otherwise, if $(\gamma-2)d_{\min}\|x^*_{T_0^c}\|_1 + \|\widetilde{A}x^*_{T_0^c}\|_2^2 \leq (\gamma+2)d_{\max}\|h_{T_0}\|_1$, then we have $\|h_{T_0^c}\|_1 \leq 2\rho\|h_{T_0}\|$; thus we may leverage Assumption RE($2s, 2\rho$) (with parameter $\delta_{2s,2\rho}$) as follows: Note that

$$\|h\|_1 \leq \|h_{T_0}\|_1 + \|h_{T_0^c}\|_1 \leq (1+2\rho)\|h_{T_0}\|_1 \leq (1+2\rho)\sqrt{s}\|h_{T_0}\|_2 \leq \frac{(1+2\rho)\sqrt{s}}{\sqrt{1-\delta_{2s,2\rho}}}\|\widetilde{A}h\|_2. \tag{A.26}$$

Furthermore, by Theorem A.1, we have

$$\|\widetilde{A}h\|_2^2 \leq 8\gamma^2 \left(\|\widetilde{A}(x^* - x_s^*)\|_2^2 + \frac{\sum_{k\in T_0} d_k^2}{1-\delta_{2s,2\rho}} \right).$$

Thus

$$\|h\|_1 \leq \frac{2\sqrt{2}\gamma(1+2\rho)\sqrt{s}}{1-\delta_{2s,2\rho}} \left((1-\delta_{2s,2\rho})\|\widetilde{A}(x^* - x_s^*)\|_2^2 + \sum_{k\in T_0} d_k^2 \right)^{1/2}. \tag{A.27}$$

For the ℓ_2 recovery error bound,

note that the k largest magnitude elements of $h_{T_0^c}$ satisfy $|h_{T_0^c}|_{(k)} \le \|h_{T_0^c}\|_1/k$, yielding

$$\|h_{T_{01}{}^c}\|_2^2 \le \|h_{T_0^c}\|_1^2 \sum_{k\ge s+1} \frac{1}{k^2} \le \frac{1}{s}\|h_{T_0^c}\|_1^2 \le \frac{4\rho^2}{s}\|h_{T_0}\|_1^2 \le 4\rho^2\|h_{T_0}\|_2^2 \le 4\rho^2\|h_{T_{01}}\|_2^2. \tag{A.28}$$

Thus,

$$\|h_{T_{01}{}^c}\|_2 \le 2\rho\|h_{T_{01}}\|_2$$

and

$$\|h\|_2 \le \|h_{T_{01}}\|_2 + \|h_{T_{01}{}^c}\|_2 \le (1+2\rho)\|h_{T_{01}}\|_2 \le \frac{1+2\rho}{\sqrt{1-\delta_{2s,2\rho}}}\|\widetilde{A}h\|_2 \tag{A.29}$$

$$\le \frac{2\sqrt{2}\gamma(1+2\rho)}{1-\delta_{2s,2\rho}} \left((1-\delta_{2s,2\rho})\|\widetilde{A}(x^* - x_s^*)\|_2^2 + \sum_{k\in T_0} d_k^2 \right)^{1/2}. \tag{A.30}$$

□

A.2 Proof of Proposition 1

Let x, J and c_0 be as in the statement of Proposition 1. We denote $s = |J|$, so we have

$$\|\widetilde{A}x\|_2^2 \ge \|\widetilde{A}x_J\|^2 + 2x_J^\top \widetilde{G} x_{J^c}.$$

But

$$\|\widetilde{A}x_J\|^2 - \|x_J\|_2^2 = x_J^\top [\widetilde{G} - I_p] x_J.$$

Hence by Assumption G(ξ),

$$|\|\widetilde{A}x_J\|^2 - \|x_J\|_2^2| \le \xi\|x_J\|_1^2 \le \xi s\|x_J\|_2^2.$$

On the other hand

$$|2x_J^\top \widetilde{G} x_{J^c}| \le 2 \sum_{k\in J, k'\in J^c} |x_k \widetilde{G}_{k,k'} x_{k'}| \le 2\xi\|x_J\|_1\|x_{J^c}\|_1 \le 2\xi c_0\|x_J\|_1^2 \le 2\xi c_0 s\|x_J\|_2^2,$$

which gives the result.

A.3 Proof of Theorem 2

We first prove the following lemma.

Lemma A.3. *Under the conditions of Theorem 2,*

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}^{(S^*)})|_k < \frac{\gamma d_k}{2} \qquad \forall k \notin S^*$$

where $\widehat{x}^{(S^*)}$ is defined in (A.6).

Proof of Lemma A.3. We have from (A.10) with $x = x^*$,

$$\|\widetilde{A}\widehat{x}^{(S^*)} - \widetilde{A}x^*\|_2^2 \le 2\gamma \sum_{k\in S^*} d_k |\widehat{x}_k^{(S^*)} - x_k^*| \tag{A.31}$$

$$\le 2\gamma \left(\sum_{k\in S^*} d_k^2 \right)^{1/2} \left(\sum_{k\in S^*} |\widehat{x}_k^{(S^*)} - x_k^*|^2 \right)^{1/2}. \tag{A.32}$$

Assumption RE applied to $\hat{x}^{(S^*)} - x^*$ gives

$$\sum_{k\in S^*} |\widehat{x}_k^{(S^*)} - x_k^*|^2 \leq (1-\delta_{s,0})^{-1} \|\widetilde{A}\widehat{x}^{(S^*)} - \widetilde{A}x^*\|_2^2. \tag{A.33}$$

Thus,

$$\|\widetilde{A}\widehat{x}^{(S^*)} - \widetilde{A}x^*\|_2^2 \leq \frac{4\gamma^2}{(1-\delta_{s,0})} \sum_{k\in S^*} d_k^2.$$

This implies by (A.33)

$$\|\widehat{x}^{(S^*)} - x^*\|_2^2 \leq \frac{4\gamma^2}{(1-\delta_{s,0})^2} \sum_{k\in S^*} d_k^2$$

and

$$\|\widehat{x}^{(S^*)} - x^*\|_1 \leq s^{1/2}\|\widehat{x}^{(S^*)} - x^*\|_2 \leq \frac{2\gamma s^{1/2}}{1-\delta_{s,0}} \left(\sum_{k\in S^*} d_k^2\right)^{1/2}.$$

Since $x^* - \widehat{x}^{(S^*)}$ is supported by S^*,

$$|\widetilde{G}(x^* - \widehat{x}^{(S^*)})|_k = \left|\sum_{\ell\in S^*} \widetilde{G}_{k,\ell}(x^* - \widehat{x}^{(S^*)})_\ell\right|.$$

But by Assumption G(ξ), if $k \neq \ell$,

$$|\widetilde{G}_{k,\ell}| \leq \xi.$$

Therefore, for all $k \notin S^*$,

$$|\widetilde{A}^\top(\widetilde{A}x^* - \widetilde{A}\widehat{x}^{(S^*)})|_k = |\widetilde{G}(x^* - \widehat{x}^{(S^*)})|_k \leq \xi\|x^* - \widehat{x}^{(S^*)}\|_1 \leq \xi\frac{2\gamma}{1-\delta_{s,0}} \left(s\sum_{k\in S^*} d_k^2\right)^{1/2}.$$

For all k in $\{0,\ldots,p-1\}$,

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

Therefore, for $k \notin S^*$,

$$\begin{aligned}|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}^{(S^*)})|_k \leq& |\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}x^*)|_k + |\widetilde{A}^\top(\widetilde{A}x^* - \widetilde{A}\widehat{x}^{(S^*)})|_k \\ \leq& d_k + \xi\frac{2\gamma}{1-\delta_{s,0}} \left(s\sum_{k\in S^*} d_k^2\right)^{1/2}.\end{aligned}$$

This in turn is bounded above by $\frac{\gamma d_k}{2}$ as soon as

$$\xi\frac{2\gamma}{1-\delta_{s,0}} \left(s\sum_{k\in S^*} d_k^2\right)^{1/2} < \left(\frac{\gamma}{2} - 1\right) \min_{k\notin S^*} d_k.$$

□

Lemma A.4. *For all $a \in \mathbb{R}^p$, the function C satisfies*

$$C(\widehat{x} + a) - C(\widehat{x}) \geq \|\widetilde{A}a\|_2^2.$$

Proof of Lemma A.4. We use Lemma A.1. For any $a \in \mathbb{R}^p$,

$$C(\widehat{x}+a) - C(\widehat{x}) = \|\widetilde{Y} - \widetilde{A}(\widehat{x}+a)\|_2^2 - \|\widetilde{Y} - \widetilde{A}\widehat{x}\|_2^2 + \gamma \sum_{k=1}^{p} d_k(|\widehat{x}_k + a_k| - |\widehat{x}_k|) \tag{A.34}$$

$$= \|\widetilde{A}a\|_2^2 - 2\langle \widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}), a\rangle + \gamma \sum_{k=1}^{p} d_k(|\widehat{x}_k + a_k| - |\widehat{x}_k|) \tag{A.35}$$

$$= \|\widetilde{A}a\|_2^2 + \gamma \sum_{k=1}^{p} d_k(|\widehat{x}_k + a_k| - |\widehat{x}_k| - a_k s_k) \geq \|\widetilde{A}a\|_2^2, \tag{A.36}$$

with $s_k = \operatorname{sign}(\widehat{x}_k)$ if $\widehat{x}_k \neq 0$ and $|s_k| \leq 1$ otherwise. The last step comes from enumeration of the possible signs of all the terms in (A.36). □

Proof of Theorem 2. We follow the approach of the proof of Proposition 2.3 of [6]. We denote $s^* = |S^*|$ and

$$S^* = \{i_1, \ldots, i_{s^*}\}.$$

We define a cost function associated with the optimization problem in (A.6), where the optimization is unconstrained over all vectors $z \in \mathbb{R}^{|S^*|}$:

$$C^{(S^*)}(z) := \|\widetilde{Y} - \widetilde{A}^{(S^*)}z\|_2^2 + \gamma \sum_{\ell=1}^{s^*} d_{i_\ell}|z_\ell|. \tag{A.37}$$

Here $\widetilde{A}^{(S^*)}$ corresponds to the submatrix of $\widetilde{A}$ such that for any $\ell \in \{1, \ldots, s^*\}$, the ℓ^{th} column of $\widetilde{A}^{(S^*)}$ is the i_ℓ^{th} column of $\widetilde{A}$. Since $1 - \delta_{s,0} > 0$, the minimizer of $C^{(S^*)}$ is unique by Lemma A.4 and Assumption RE. Therefore, if $\widehat{z} = \operatorname{argmin}_z C^{(S^*)}(z)$, then for any $\ell \in \{1, \ldots, s^*\}$, $\widehat{z}_\ell = \widehat{x}_{i_\ell}^{(S^*)}$. The first-order optimality condition for $C^{(S^*)}(z)$ implies that for all $\ell \in \{1, \ldots, s^*\}$,

$$((\widetilde{A}^{(S^*)})^\top(\widetilde{Y} - \widetilde{A}^{(S^*)}\widehat{z}))_\ell = \frac{\gamma d_{i_\ell}}{2}\operatorname{sign}(\hat{z}_\ell) \qquad \text{for } \ell \text{ s.t. } \hat{z}_\ell \neq 0 \tag{A.38}$$

$$|(\widetilde{A}^{(S^*)})^\top(\widetilde{Y} - \widetilde{A}^{(S^*)}\widehat{z})|_\ell \leq \frac{\gamma d_{i_\ell}}{2} \qquad \text{for } \ell \text{ s.t. } \hat{z}_\ell = 0. \tag{A.39}$$

or, equivalently

$$(\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}^{(S^*)}))_k = \frac{\gamma d_k}{2}\operatorname{sign}(\widehat{x}_k^{(S^*)}) \qquad \text{for } k \in S^* \text{ s.t. } \widehat{x}_k^{(S^*)} \neq 0 \tag{A.40}$$

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}^{(S^*)})|_k \leq \frac{\gamma d_k}{2} \qquad \text{for } k \in S^* \text{ s.t. } \widehat{x}_k^{(S^*)} = 0. \tag{A.41}$$

For $k \notin S^*$, by Lemma A.3 we have

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}^{(S^*)})|_k < \frac{\gamma d_k}{2}.$$

Thus, $\widehat{x}^{(S^*)}$ satisfies the first-order optimality condition for C and hence $\widehat{x}^{(S^*)}$ minimizes $C(x)$ over all $x \in \mathbb{R}^p$. This in turn implies

$$0 = |C(\widehat{x}) - C(\widehat{x}^{(S^*)})| \geq \|\widetilde{A}(\widehat{x} - \widehat{x}^{(S^*)})\|_2^2 \geq 0$$

where the middle inequality comes from Lemma A.4. We thus have

$$\widetilde{A}\widehat{x} = \widetilde{A}\widehat{x}^{(S^*)}$$

and so for all $k \notin S^*$

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x}^{(S^*)})|_k < \frac{\gamma d_k}{2} \tag{A.42}$$

implies

$$|\widetilde{A}^\top(\widetilde{Y} - \widetilde{A}\widehat{x})|_k < \frac{\gamma d_k}{2}. \tag{A.43}$$

We then have that $\mathrm{supp}(\widehat{x}) \subset S^*$ if (2.8) is true.

□

A.4 Proof of Proposition 2

Proof of Proposition 2. 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 $k \in S^*$,

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

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

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

Finally, using Assumption RE, we obtain

$$\begin{aligned}\|\widehat{x}^{\mathrm{LS}} - x^*\|_2^2 \leq& (1-\delta_{s,0})^{-2}\|\widetilde{A}_{S^*}^H(\widetilde{Y} - \widetilde{A}x^*)\|_2^2 \\ \leq& (1-\delta_{s,0})^{-2}\sum_{k\in S^*} d_k^2.\end{aligned}$$

□

A.5 Proof of Theorem 3

Proof of Theorem 3. We have from Lemma A.2, with $x = x^*$,

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq 2\gamma \sum_{k\in S^*} d_k|\widehat{x}_k - x_k^*| \tag{A.44}$$

$$\leq 2\gamma \left(\sum_{k\in S^*} d_k^2\right)^{1/2}\left(\sum_{k\in S^*}|\widehat{x}_k - x_k^*|^2\right)^{1/2}. \tag{A.45}$$

Since $\mathrm{supp}(\widehat{x}) \subset S^*$, Assumption RE gives

$$\left(\sum_{k\in S^*}|\widehat{x}_k - x_k^*|^2\right)^{1/2} \leq (1-\delta_{s,0})^{-1/2}\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2.$$

Thus,

$$\|\widetilde{A}\widehat{x} - \widetilde{A}x^*\|_2^2 \leq \frac{4\gamma^2}{1-\delta_{s,0}}\sum_{k\in S^*} d_k^2.$$

This implies

$$\|\widehat{x} - x^*\|_2^2 \leq \frac{4\gamma^2}{(1-\delta_{s,0})^2} \sum_{k \in S^*} d_k^2$$

and

$$\|\widehat{x} - x^*\|_1 \leq s^{1/2} \|\widehat{x} - x^*\|_2 \leq \frac{2\gamma s^{1/2}}{1-\delta_{s,0}} \left(\sum_{k \in S^*} d_k^2 \right)^{1/2}.$$

We next obtain the the sup-norm bound. By Lemma A.1, for any k,

$$|\widetilde{A}^\top (\widetilde{Y} - \widetilde{A}\widehat{x})|_k \leq \frac{\gamma d_k}{2}. \tag{A.46}$$

From here we have

$$\begin{aligned}
\|x^* - \widehat{x}\|_\infty =& \left\| \left(\widetilde{A}^\top \widetilde{A} - \widetilde{A}^\top \widetilde{A} + I_p \right) (x^* - \widehat{x}) \right\|_\infty && \text{(A.47)} \\
\leq& \left\| \widetilde{A}^\top \widetilde{A} (x^* - \widehat{x}) \right\|_\infty + \left\| \left(\widetilde{A}^\top \widetilde{A} - I_p \right) (x^* - \widehat{x}) \right\|_\infty && \text{(A.48)} \\
=& \|\widetilde{A}^\top (\widetilde{Y} - \widetilde{A}x^*) - \widetilde{A}^\top (\widetilde{Y} - \widetilde{A}\widehat{x})\|_\infty + \left\| \left(\widetilde{A}^\top \widetilde{A} - I_p \right) (x^* - \widehat{x}) \right\|_\infty && \text{(A.49)} \\
\leq& d_{\max}(1 + \gamma/2) + \xi \|x^* - \widehat{x}\|_1 && \text{(A.50)} \\
\leq& d_{\max}(1 + \gamma/2) + \frac{2\xi\gamma s^{1/2}}{1-\delta_{s,0}} \left(\sum_{k \in S^*} d_k^2 \right)^{1/2} && \text{(A.51)} \\
\leq& d_{\max}(1 + \gamma/2) + (\frac{\gamma}{2} - 1) \min_{k \notin S^*} d_k && \text{(A.52)} \\
\leq& \gamma d_{\max}. && \text{(A.53)}
\end{aligned}$$

where (A.50) follows from Lemma A.1 and Assumptions Weights($\{d_k\}_k$) and G(ξ), (A.51) follows from (2.10) and (A.52) follows from (2.8). □

B Concentration inequality for data-dependent weights (proof of Lemma 1)

The proof of (3.2) is really classical and follows the lines of Bernstein inequality. Let $Z = R^\top Y$ and $z = R^\top A x^*$. 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)

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

If $\lambda < (3/b)$, then by classical computations (see [29] 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 $\lambda < (3/b)$,

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

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 [29]).

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 $-R_2$ and for (3.5), to combine both (3.3) and (3.4).

C Validation of assumptions for Bernoulli sensing of Section 4

In the sequel, the notation $\square$ represents an absolute constant that may change from line to line.

C.1 Rescaling and recentering

First let us prove that $\mathbb{E}(\tilde{G}) = I_p$, with $\tilde{G} = \widetilde{A}^\top \widetilde{A}$.

Indeed, the (k, k') element of $\tilde{G}$ is

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

Hence $\mathbb{E}(\tilde{G}_{k,k'}) = 0$ if $k \neq k'$ and $\mathbb{E}(\tilde{G}_{k,k}) = 1$. 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

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

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 G holds (proof of Proposition 3)

Proof of Proposition 3. We have for any k, ℓ,

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

Therefore, $\mathbb{E}[\widetilde{G}_{k\ell}] = 1_{k=\ell}$. We apply Bernstein concentration inequality to

$$X_i = \frac{(a_{ik} - q)(a_{i\ell} - q)}{q(1-q)},$$

which are independent. Since, on the one hand

$$|X_i| \leq b := \max\left(\frac{1-q}{q}, \frac{q}{1-q}\right),$$

and on the other hand, by setting

$$v_n = \sum_{i=1}^{n} \mathbb{E}\left[\frac{(a_{ik} - q)^2 (a_{i\ell} - q)^2}{q^2(1-q)^2}\right]$$

we have for $k \neq \ell$, $v_n = n$ and for $k = \ell$,

$$v_n = \sum_{i=1}^{n} \mathbb{E}\left[\frac{(a_{ik} - q)^2 (a_{i\ell} - q)^2}{q^2(1-q)^2}\right] = n\mathbb{E}\left[\frac{(a_{1k} - q)^4}{q^2(1-q)^2}\right] = n\left[\frac{(1-q)^2}{q} + \frac{q^2}{1-q}\right]$$

with probability larger than $1 - 2p^2 e^{-\theta}$,

$$|n(\widetilde{G} - I_p)_{k\ell}| \leq \sqrt{2v_n\theta} + \frac{b\theta}{3}.$$

So, with probability larger than $1 - 2p^2 e^{-\theta}$,

$$|(\widetilde{G} - I_p)_{k\ell}| \leq \xi,$$

with

$$\xi = \sqrt{\frac{2\theta}{n}\left(\frac{(1-q)^2}{q} + \frac{q^2}{1-q}\right)} + \frac{\theta}{3n} \max\left(\frac{1-q}{q}, \frac{q}{1-q}\right).$$

The proposition follows from setting $\theta = 3 \log p$. □

C.3 Proofs for data-dependent weights

Proposition 5. *Let*

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

with

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

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

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

where $\hat{N}$ is an estimator of $\|x^*\|_1$ given by

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

satisfies Assumption Weights(d).

Proposition 6. *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 7. *With the same notations and assumptions as Proposition 5, there exists absolute constants c, c' such that with probability larger than $1-c'/p$, the choice (depending on k)*

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

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

with the vector V_k of size n given by

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

satisfies Assumption Weights($\{d_k\}_k$).

Proposition 8. *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 5 and 7)

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 5, we use a bound on $\|T_1\|_\infty + \|T_2\|_\infty$. To derive the non-constant weights of Proposition 7, 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 [21]. 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 choice of functions c_ℓ, $b_{\ell'}$ such that $\mathbb{E}(\sum_{\ell=2}^n c_\ell(a_{\ell,k})^2) \leq 1$ and $\mathbb{E}(\sum_{\ell'=1}^{n-1} b_{\ell'}(a_{\ell',k'})^2) \leq 1$. But

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

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. For any $\theta > 0$, the concentration inequality of [21] involves up to absolute multiplicative constants, a term of the form $C_{\mathbb{M}}\sqrt{\theta} + D_{\mathbb{M}}\theta + B_{\mathbb{M}}\theta^{3/2} + A_{\mathbb{M}}\theta^2$ in which the main terms are $D_{\mathbb{M}}\theta$ and $A_{\mathbb{M}}\theta^2$ by the previous computations. Therefore there exists some absolute constants c_1 and c_2 such that as soon as $n \geq 2$, with probability larger than $1 - c_1 p^2 e^{-\theta}$, for all k, k'

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

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 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 $1 - c_3 p^2 e^{-\theta}$,

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

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 $\ell = 1, \ldots, n$,

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

By (3.3) of Lemma 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

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

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 $\theta = 3\log(p)$ gives Proposition 5. On the other hand, one could have applied (3.5) of Lemma 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 7.

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

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

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

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

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

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

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

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

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

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 weights (proof of Proposition 6)

Applying (3.2) of Lemma 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 $1 - \square p e^{-\theta}$

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

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

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

But by the previous computations, W is of the order of $\frac{1}{n\min(q,1-q)}$, which gives Proposition 6 with $\theta = 3\log(p)$.

C.3.4 Control of the non-constant weights (proof of Proposition 8)

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

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

But $V_k^\top Ax^* = \sum_{u=1}^p w(u,k) x_u^*$ 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 8 with $\theta = 3\log(p)$.

D Validation of assumptions for random convolution model of Section 5

D.1 Rescaling and recentering

Note that Proposition 9 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 $\mathbb{E}(\widetilde{A}^{\mathsf{H}}(\widetilde{Y} - \widetilde{A}x^*)) = 0$ as expected.

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

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

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:

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

□

D.2 Assumption G holds (proof of Proposition 4)

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 9. *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 $1 - 5.54\, p e^{-\theta}$ 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 actually that Assumption G($\xi(\theta)$) is satisfied on the event $\Omega_{\mathbb{U}}(\theta)$. Proposition 4 follows with $\theta = 2 \log p$.

Proof of Proposition 9. Let $\beta_0 = \frac{1}{\sqrt{m}}$ and $\beta_1 = \frac{\sqrt{m}-1}{p}$. For all $k \neq \ell \in \{0, \dots, p-1\}$,

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

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

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 [21]. 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 $a \in \{0, \dots, p-1\}$,

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

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,

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

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 [21], 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 [21, 17]. 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 10. *The constant weights given by (5.4) satisfy Assumption Weights(d) with probability larger than $1-C/p$ for some absolute positive constant C.*

Proposition 11. *Under the notations of Proposition 10, there exists positive absolute constants c and C and an event of probability larger than $1-C/p$ such that on this event*

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

Proposition 12. *The non-constant weights given by (5.5) satisfy Assumption Weights(d) with probability larger than $1-C/p$ for some absolute positive constant C.*

Proposition 13. *Under the notations of Proposition 10, there exists some absolute constants $\kappa_1,\kappa_2,c_1,c_2$ and C positive such that if $p\ge 5$ and if*

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

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

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

Thanks to those upper and lower bounds on the d_k's it is easy to see that (2.8) is matched as soon as

$$\xi^2\frac{4\gamma^2}{(1-s\xi)^2}\left[s\frac{\|x^*\|_1\log(p)}{m}+s^2\frac{\|x^*\|_1\log(p)^2}{p}+\frac{s^2\log(p)^4}{m^2}\right]\le\square\left(\frac{\gamma}{2}-1\right)^2\left[\|x^*\|_1\frac{\log(p)}{p}+\frac{\log(p)^2}{m^2}\right].$$

Since $\xi\simeq\frac{\log p}{\sqrt{p}}$ under (D.9), which is implied by (5.6), as soon as $s\xi<1/2$ this is implied by

$$\frac{\log(p)^2}{p}\left(s\frac{\log(p)}{m}+s^2\frac{\log(p)^2}{p}\right)\le\square_\gamma\frac{\log(p)}{p},$$

and

$$\frac{\log(p)^2}{p}\frac{s^2\log(p)^4}{m^2}\le\square_\gamma\frac{\log(p)^2}{m^2}.$$

Both of them are implied by

$$s\lesssim\sqrt{p}\log(p)^{-2}.$$

D.3.1 Assumption Weights holds (proof of Propositions 10 and 12)

Proposition 12 is just the application of (3.5) of Lemma 1 to each of the vectors R_k with $\theta = 2\log(p)$.

For Proposition 10 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

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

But one could apply Lemma 1 with $R = -\mathbb{1}$ to obtain that

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

which is equivalent to

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

Therefore combining (3.3) of Lemma 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 $w(\ell)$s

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 $\{0, \dots, p-1\}$,*

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

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 $1 - pe^{-\theta}$,*

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

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

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

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

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

But $1/(p^2 m) \le \min(\theta/p, \theta^2/m)$, 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 9 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 weights (proof of Proposition 11)

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 13)

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

$$\hat{v}_k \leq \square \left[v_k + B^2\theta \right].$$

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

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

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

For the lower bound, the arguments are similar

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

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

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

which gives the result.