Chapter 2
Preliminaries

2.1 Sparse Linear Regression and Compressed Sensing

Least squares problems occur in various signal processing and statistical inference applications. In these problems the relation between the vector of noisy observations $\mathbf{y} \in \mathbb{R}^m$ and the unknown parameter or signal $\mathbf{x}^\star \in \mathbb{R}^n$ is governed by a linear equation of the form

$$\mathbf{y} = \mathbf{A}\mathbf{x}^\star + \mathbf{e}, \tag{2.1}$$

where $\mathbf{A} \in \mathbb{R}^{m \times n}$ is a matrix that may model a linear system or simply contain a set of collected data. The vector $\mathbf{e} \in \mathbb{R}^m$ represents the additive observation noise. Estimating $\mathbf{x}^\star$ from the observation vector $\mathbf{y}$ is achieved by finding the vector $\mathbf{x}$ that minimizes the squared error $\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2{}^2$. This least squares approach, however, is well-posed only if the nullspace of matrix $\mathbf{A}$ merely contains the zero vector. The cases in which the nullspace is greater than the singleton $\{\mathbf{0}\}$, as in *underdetermined* scenarios ($m < n$), are more relevant in a variety of applications. To enforce unique least squares solutions in these cases, it becomes necessary to have some prior information about the structure of $\mathbf{x}^\star$.

One of the structural characteristics that describe parameters and signals of interest in a wide range of applications from medical imaging to astronomy is *sparsity*. Study of high-dimensional linear inference problems with sparse parameters has gained significant attention since the introduction of Compressed Sensing, also known as *Compressive Sampling*, (CS) Donoho (2006); Candès and Tao (2006). In standard CS problems the aim is to estimate a sparse vector $\mathbf{x}^\star$ from linear measurements. In the absence of noise (i.e., when $\mathbf{e} = \mathbf{0}$), $\mathbf{x}^\star$ can be determined uniquely from the observation vector $\mathbf{y} = \mathbf{A}\mathbf{x}^\star$ provided that spark $(\mathbf{A}) > 2\|\mathbf{x}^\star\|_0$ (i.e., every $2\|\mathbf{x}^\star\|_0$ columns of $\mathbf{A}$ are linearly independent) Donoho and Elad (2003). Then the ideal estimation procedure would be to find the sparsest vector $\mathbf{x}$ that

S. Bahmani, *Algorithms for Sparsity-Constrained Optimization*, Springer Theses 261,
DOI 10.1007/978-3-319-01881-2_2, © Springer International Publishing Switzerland 2014

incurs no residual error (i.e., $\|\mathbf{Ax}-\mathbf{y}\|_2 = 0$). This ideal estimation method can be extended to the case of noisy observations as well. Formally, the vector $\mathbf{x}^\star$ can be estimated by solving the ℓ_0-minimization

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t. } \|\mathbf{y}-\mathbf{Ax}\|_2 \leq \varepsilon, \tag{2.2}$$

where ε is a given upper bound for $\|\mathbf{e}\|_2$ Candès et al. (2006). Unfortunately, the ideal solver (2.2) is computationally NP-hard in general Natarajan (1995) and one must seek approximate solvers instead.

It is shown in Candès et al. (2006) that under certain conditions, minimizing the ℓ_1-norm as a convex proxy for the ℓ_0-norm yields accurate estimates of $\mathbf{x}^\star$. The resulting approximate solver basically returns the solution to the convex optimization problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t. } \|\mathbf{y}-\mathbf{Ax}\|_2 \leq \varepsilon, \tag{2.3}$$

The required conditions for approximate equivalence of (2.2) and (2.3), however, generally hold only if measurements are collected at a higher rate. Ideally, one merely needs $m = O(s)$ measurements to estimate $\mathbf{x}^\star$, but $m = O(s \log n/s)$ measurements are necessary for the accuracy of (2.3) to be guaranteed.

The convex program (2.3) can be solved in polynomial time using interior point methods. However, these methods do not scale well as the size of the problem grows. Therefore, several first-order convex optimization methods are developed and analyzed as more efficient alternatives (see, e.g., Figueiredo et al. 2007; Hale et al. 2008; Beck and Teboulle 2009; Wen et al. 2010; Agarwal et al. 2010). Another category of low-complexity algorithms in CS are the non-convex *greedy pursuits* including Orthogonal Matching Pursuit (OMP) Pati et al. (1993); Tropp and Gilbert (2007), Compressive Sampling Matching Pursuit (CoSaMP) Needell and Tropp (2009), Iterative Hard Thresholding (IHT) Blumensath and Davies (2009), and Subspace Pursuit Dai and Milenkovic (2009) to name a few. These greedy algorithms implicitly approximate the solution to the ℓ_0-constrained least squares problem

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{y}-\mathbf{Ax}\|_2^2 \quad \text{s.t. } \|\mathbf{x}\|_0 \leq s. \tag{2.4}$$

The main theme of these iterative algorithms is to use the residual error from the previous iteration to successively approximate the position of non-zero entries and estimate their values. These algorithms have shown to exhibit accuracy guarantees similar to those of convex optimization methods, though with more stringent requirements.

As mentioned above, to guarantee accuracy of the CS algorithms the measurement matrix should meet certain conditions such as *incoherence* Donoho and Huo (2001), Restricted Isometry Property (RIP) Candès et al. (2006), Nullspace Property

Cohen et al. (2009), etc. Among these conditions RIP is the most commonly used and the best understood condition. Matrix $\mathbf{A}$ is said to satisfy the RIP of order k—in its symmetric form—with constant δ_k, if $\delta_k < 1$ is the smallest number that

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

holds for all k-sparse vectors $\mathbf{x}$. Several CS algorithms are shown to produce accurate solutions provided that the measurement matrix has a sufficiently small RIP constant of order ck with c being a small integer. For example, solving (2.3) is guaranteed to yield an accurate estimate of s-sparse $\mathbf{x}^\star$ if $\delta_{2s} < \sqrt{2} - 1$ Candès (2008). Interested readers can find the best known RIP-based accuracy guarantees for some of the CS algorithms in Foucart (2012).

The formulation of sparse linear regression problems as well as algorithms used to solve them are virtually identical to CS. However, these problems that are usually studied in statistics and machine learning, have a set-up that distinguishes them from the CS problems. The sensing or sampling problems addressed by CS often do not impose strong restrictions on the choice of the measurement matrix. Matrices drawn from certain ensembles of random matrices (e.g., Gaussian, Rademacher, partial Fourier, etc.) can be chosen as the measurement matrix Candès and Tao (2006). These types of random matrices allow us to guarantee the required conditions such as RIP, at least in the probabilistic sense. However, the analog of the measurement matrix in sparse linear regression, the *design matrix,* is often dictated by the data under study. In general the entries of the design matrix have unknown distributions and are possibly dependent. In certain scenarios the independence of observations/measurements may not hold either. While it is inevitable to make assumptions about the design matrix for the purpose of theoretical analysis, the considered assumptions are usually weaker compared to the CS assumptions. Consequently, the analysis of sparse linear inference problems is more challenging than in CS problems.

2.2 Nonlinear Inference Problems

To motivate the need for generalization of CS, in this section we describe a few problems and models which involve non-linear observations.

2.2.1 Generalized Linear Models

Generalized Linear Models (GLMs) are among the most commonly used models for parametric estimation in statistics Dobson and Barnett (2008). Linear, logistic, Poisson, and gamma models used in corresponding regression problems all belong to the family of GLMs. Because the parameter and the data samples in GLMs

are mixed in a linear form, these models are considered among linear models in statistics and machine learning literature. However, as will be seen below, in GLMs the relation between the response variable and the parameters is in general nonlinear.

Given a vector of covariates (i.e., data sample) $\mathbf{a} \in \mathcal{X} \subseteq \mathbb{R}^n$ and a true parameter $\mathbf{x}^\star \in \mathbb{R}^n$, the response variable $y \in \mathcal{Y} \subseteq \mathbb{R}$ in canonical GLMs is assumed to follow an exponential family conditional distribution: $y \mid \mathbf{a}; \mathbf{x}^\star \sim Z(y) \exp\left(y \langle \mathbf{a}, \mathbf{x}^\star \rangle - \psi\left(\langle \mathbf{a}, \mathbf{x}^\star \rangle\right)\right)$, where $Z(y)$ is a positive function, and $\psi : \mathbb{R} \mapsto \mathbb{R}$ is the *log-partition function* that satisfies $\psi(t) = \log \int_{\mathcal{Y}} Z(y) \exp(ty) \, \mathrm{d}y$ for all $t \in \mathbb{R}$. Examples of the log-partition function, which is always convex, include but are not limited to $\psi_{\text{lin}}(t) = t^2/2\sigma^2$, $\psi_{\log}(t) = \log(1 + \exp(t))$, and $\psi_{\text{Pois}}(t) = \exp(t)$ corresponding to linear, logistic, and Poisson models, respectively.

Suppose that m iid covariate-response pairs $\{(\mathbf{a}_i, y_i)\}_{i=1}^m$ are observed in a GLM. As usual, it is assumed that $\mathbf{a}_i$'s do not depend on the true parameter. The joint likelihood function of the observation at parameter $\mathbf{x}$ can be written as $\prod_{i=1}^m p(\mathbf{a}_i) \, p(y_i \mid \mathbf{a}_i; \mathbf{x})$ where $p(y_i \mid \mathbf{a}_i; \mathbf{x})$ is the exponential family distribution mentioned above. In the Maximum Likelihood Estimation (MLE) framework the negative log likelihood is used as a measure of the discrepancy between the true parameter $\mathbf{x}^\star$ and an estimate $\mathbf{x}$ based on the observations. Because $p(\mathbf{a}_i)$'s do not depend on $\mathbf{x}$ the corresponding terms can be simply ignored. Formally, the average of negative log conditional likelihoods is considered as the empirical loss

$$f(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \psi\left(\langle \mathbf{a}_i, \mathbf{x} \rangle\right) - y_i \langle \mathbf{a}_i, \mathbf{x} \rangle ,$$

and the MLE is performed by minimizing $f(\mathbf{x})$ over the set of feasible $\mathbf{x}$. The constant c and $Z(y)$ that appear in the distribution are disregarded as they have no effect in the outcome. We will use the logistic model, a special case of GLMs, in Chaps. 3 and 5 as examples where our algorithms apply.

2.2.2 1-Bit Compressed Sensing

As mentioned above, the ideal CS formulation allows accurate estimation of sparse signals from a relatively small number of linear measurements. However, sometimes certain practical limitations impose non-ideal conditions that must be addressed in order to apply the CS framework. One of these limitations is the fact that in digital signal processing systems the signals and measurements have quantized values. Motivated by this problem, researchers have studied the performance of CS with quantized measurements. Of particular interest has been the problem of 1-bit Compressed Sensing Boufounos and Baraniuk (2008), in which the CS linear measurements are quantized down to one bit that represents their sign. Namely, for a signal $\mathbf{x}^\star$ and measurement vector $\mathbf{a}$ the observed measurement in 1-bit CS is

given by $y = \text{sgn}(\langle \mathbf{a}, \mathbf{x}^\star \rangle + e)$ where e is an additive noise. As can be seen, the observations and the signal are related by a nonlinear transform. In Chap. 4 we will explain how the problem of estimating $\mathbf{x}^\star$ from a collection of 1-bit measurements can be cast as a sparsity-constrained optimization.

2.2.3 Phase Retrieval

One of the common non-linear inverse problems that arise in applications such as optics and imaging is the problem of *phase retrieval*. In these applications the observations of the object of interest are in the form of phaseless linear measurements. In general, reconstruction of the signal is not possible in these scenarios. However, if the signal is known to be sparse *a priori* then accurate reconstruction can be achieved up to a unit-modulus factor. In particular, *Quadratic Compressed Sensing* is studied in Shechtman et al. (2011b,a) for phase retrieval problems in sub-wavelength imaging. Using convex relaxation it is shown that the estimator can be formulated as a solution to a Semi-Definite Program (SDP) dubbed *PhaseLift* Candès et al. (2012); Candès and Li (2012); Li and Voroninski (2012).

References

A. Agarwal, S. Negahban, and M. Wainwright. Fast global convergence rates of gradient methods for high-dimensional statistical recovery. In J. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23, pages 37–45. 2010. long version available at arXiv:1104.4824v1 [stat.ML].

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, Nov. 2009.

P. Boufounos and R. Baraniuk. 1-bit compressive sensing. In *Information Sciences and Systems, 2008. CISS 2008. 42nd Annual Conference on*, pages 16–21, Mar. 2008.

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

E. J. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. arXiv:1208.6247 [cs.IT], Aug. 2012.

E. J. Candès and T. Tao. Near optimal signal recovery from random projections: universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, Dec. 2006.

E. J. Candès, J. K. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006.

E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 2012. DOI 10.1002/cpa.21432.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of American Mathematical Society*, 22(1):211–231, Jan. 2009.

W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *IEEE Transactions on Information Theory*, 55(5):2230–2249, 2009.

A. J. Dobson and A. Barnett. *An Introduction to Generalized Linear Models*. Chapman and Hall/CRC, Boca Raton, FL, 3rd edition, May 2008. ISBN 9781584889502.

D. L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7):2845–2862, 2001.

M. Figueiredo, R. Nowak, and S. Wright. Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems. *IEEE Journal of Selected Topics in Signal Processing*, 1(4):586–597, Dec. 2007. ISSN 1932–4553. DOI 10.1109/JSTSP.2007.910281.

S. Foucart. Sparse recovery algorithms: sufficient conditions in terms of restricted isometry constants. In *Approximation Theory XIII: San Antonio 2010*, volume 13 of *Springer Proceedings in Mathematics*, pages 65–77, San Antonio, TX, 2012. Springer New York.

E. Hale, W. Yin, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: methodology and convergence. *SIAM Journal on Optimization*, 19(3):1107–1130, 2008.

X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. arXiv:1209.4785 [cs.IT], Sept. 2012.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Conference Record of the 27th Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 40–44, Pacific Grove, CA, Nov. 1993.

Y. Shechtman, Y. C. Eldar, A. Szameit, and M. Segev. Sparsity based sub-wavelength imaging with partially incoherent light via quadratic compressed sensing. *Optics Express*, 19(16):14807–14822, July 2011a.

Y. Shechtman, A. Szameit, E. Osherovic, E. Bullkich, H. Dana, S. Gazit, S. Shoham, M. Zibulevsky, I. Yavneh, E. B. Kley, Y. C. Eldar, O. Cohen, and M. Segev. Sparsity-based single-shot sub-wavelength coherent diffractive imaging. In *Frontiers in Optics*, OSA Technical Digest, page PDPA3. Optical Society of America, Oct. 2011b.

J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

Z. Wen, W. Yin, D. Goldfarb, and Y. Zhang. A fast algorithm for sparse reconstruction based on shrinkage, subspace optimization, and continuation. *SIAM Journal on Scientific Computing*, 32(4):1832–1857, 2010.

http://www.springer.com/978-3-319-01881-2

Algorithms for Sparsity-Constrained Optimization

Bahmani, S.

2014, XXI, 107 p. 13 illus., 12 illus. in color.,
ISBN: 978-3-319-01881-2