When does Alternating Descent Conditional Gradient work better than non-convex optimization with random restarts?

Charles Zheng and ?

November 23, 2015

Abstract

Alternating Descent Conditional Gradient (ADCG) is a method for solving sparse inverse problems which combines nonconvex and convex optimization techniques. Although ADCG is observed to achieve superior performance to alternative approaches, little is known about its performance due to its non-convex subroutines. In this work, we consider a sparse inverse problem on an infinite lattice, and given a general nonlinear optimization subroutine, compare the performance of ADCG equipped with the given subroutine versus the approach of simply applying the subroutine with random intializations, supposing that the true signal is known to lie in a finite sublattice. Under a number of symmetry assumptions, we show that under an asymptotic regime where the number of sources is growing, the volume of the support of the signal is relatively small compared to the volume of the sublattice, and where the signal size is large compared to the noise level, that ADCG converges to the global minimum with fixed number of calls while the probability of reaching the global minimum under a random intitialization goes to zero.

1 Introduction

In numerous applications, one is interested in reconstructing the locations and parameters of multiple signal sources from noisy observations. For instance, in super-resolution imaging, one uses a microscope to obtain a 2D image of a number of fluorescing point sources, and the goal is to recover the locations of the point sources on the slide.

Such a sparse inverse problem is described by a known, or unknown number of sources K, a vector of parameters $\theta_i \in \mathbb{R}^p$ describing the ith source (e.g. location in space), and a positive real weight w_i giving amplitude of the signal from the ith source. The signal from a single source θ is a function in $\mathbb{R}^d$, denoted by ψ_θ, and the combined signal from all K sources is given by the function

$$\mu(x) = \sum_{i=1}^{K} w_i \psi_{\theta_i}(x).$$

In the case of super-resolution imaging, for example, the parameter $\theta \in \mathbb{R}^2$ gives the location of the point source on the slide, w_i gives the intensity of the fluorescence, and $\phi_\theta(x)$ is a symmetric kernel function centered at θ, given by the point-spread function of the lens.

Given measurement locations $x_1, \ldots, x_N$, we observe signals

$$y_i \sim F(\mu(x_i)),$$

where $\{F(\mu)\}_{\mu \in \mathbb{R}}$ is a family of parametric distributions. For instance, a simple case is $F(\mu) = N(\mu, \sigma^2)$, corresponding to the familiar Gaussian error model

$$y_i = \mu(x_i) + \epsilon_i, \ \epsilon_i \sim N(0, \sigma^2).$$

Very shortly, we will assume that $F(\mu)$ has a density $f_\mu(y)$ on $\mathbb{R}$, and hence a negative log-likelihood function

$$\ell(y; \mu) = -\log(f_\mu(y))$$

Having observed $y_1, \ldots, y_N$, our goal is to estimate the number of sources K (if K is unknown), and to recover the locations $\theta_1, \ldots, \theta_K$. A natural approach is to minimize the objective function

$$\text{minimize} \sum_{i=1}^{n} \ell\left(y_i; \sum_{j=1}^{K} w_j \psi_{\theta_j}(x_i)\right)$$

subject to some sparsity constraint, where ℓ is the log-likelihood function for $\{F(\mu)\}$. However, throughout the literuature, a variety of methods have been proposed for choosing the sparsity constraint and for minimizing the objective function. We describe three main categories of methods:

- Using nonlinear optimization (e.g. gradient descent or Newton's method) to minimize the objective function subject to a constraint on the putative number of sources K.

- Discretizing the parameter space by choosing candidate parameters $\theta_1, \ldots, \theta_m \in \mathbb{R}^p$, then finding the optimal weights

  $$\text{minimize}_w \sum_{i=1}^{n} \ell \left(y_i; \sum_{j=1}^{m} w_j \psi_{\theta_j}(x_i) \right),$$

  possibly subject to an L_1-norm constraint or penalty on the weights w. Note that the discretized problem is *convex* if ℓ is convex.

- Alternating Descent Conditional Gradient. (To be described below.)

Nonlinear optimization is not guaranteed to achieve the global minimum of the objective function, hence a usual approach is to run nonlinear optimization multiple times with random starting conditions. As a result, it is often quite costly to get a good solution of the optimization problem using nonlinear optimization.

In contrast, when ℓ is convex, it is possible to deterministically approximate the global minimum with the discretization approach, by choosing a suitably fine discretization and then applying convex optimization to solve the discretized objective function. However, one is limited to using convex constraints, which excludes the possibility of solving the optimization problem subject to a constraint or penalty on the number of sources K, which is equal to the L_0 norm of the weights. Instead, here one typically places a constraint or penalty on the L_1 norm of the weights, since the L_1-norm is the tightest convex relaxation of the L_0 norm. Intuitively, one expects the L_1 convex relaxation to yield a worse solution than the L_0 constrained problem; while plenty of theoretical results (Morgenshtern, Candes, etc.) establish the statistical properties of the estimators resulting from L_1 minimization, little is known about the comparative performance of L_0-constrained minimization, even supposing that the global minima are achieved. In particular, it is not possible to apply the sparse recovery results of Donoho et al., since the design matrix in our problem is typically highly collinear and hence violates the usual L_1 support recovery conditions. It is worth mentioning the work by Slawski (2012), which introduces a framework for studying the sparse recovery problem given such highly correlated design matrices, and which also provides results on denoising.

Alternating Descent Conditional Gradient (Boyd et al.) combines the convex and nonconvex approaches. It is shown to have guaranteed performance to the global minimizer, subject to *convex* sparsity constraints. The algorithm is defined with reference to a gradient subroutine τ and a nonlinear descent subroutine ν. The gradient subroutine is given residuals $r_1, \ldots, r_N$ as input, and and outputs the parameter θ whose signal maximizes the inner product with the gradient of the loss with respect to the residuals:

$$\tau(r_1, \ldots, r_n) = \mathrm{argmax}_\theta \sum_{i=1}^{n} \psi_\theta(x_i) \dot{\ell}(r_i).$$