Random Fields and Optimization

Charles Zheng and ?

November 25, 2015

Abstract

One of the most common approaches for minimizing a function $f(x)$ over a compact domain is to apply a deterministic local descent method, although such an approach is highly likely to yield a local minimum. Using techniques borrowed from the optimal transport literature, we propose a new method for characterizing the probability of convergence to the global minimum from a random starting point. In limit of infinitely small step sizes, the local descent iterates $x^{(0)}, \ldots,$ trace out a continuous trajectory $x(t)$, where t is the limit of the iteration number times the step size. Consequently, the domains of attraction for local minima are simply connected and partition the domain, and the probability of convergence to the global minimum is equal to the mass of the initialization distribution contained in the domains of attraction of global minima. One can therefore achieve bounds on the probability of global convergence by bounding the *radius* or *diameter* of the domain of attraction. While such computations may be generally intractable, here we obtain average-case bounds for such quantities for f drawn from a random ensemble, e.g., a Gaussian process. This produces convergence probabilities for a variety of applications: for instance, sparse inverse problems.

1 Introduction

Suppose we are given a function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X}$ is a compact domain in $\mathbb{R}^p$, and we have access to the gradient ∇f, and possibly the Hessian Hf. Then we can solve the optimization problem

$$\text{minimize} f(x)$$

by applying a local descent method, such as gradient descent or Newton's method. In general, such a method can be characterized by a direction

mapping $g : \mathcal{X} \to \mathbb{R}^p$, and produces iterates $x^{(1)}, x^{(2)}, \ldots$ given by

$$x^{(k+1)} = x^{(k)} - \epsilon g(x^{(k)}),$$

defining $x^{(0)}$ to be the starting point x_0. In stochastic descent methods, g can be a random mapping, but throughout this paper we limit our scope to deterministic methods. Two well-known examples are gradient descent,

$$g(x) = \nabla f(x)$$

and Newton's method,

$$g(x) = (Hf(x))^{-1} \nabla f(x)$$

Having chosen a step size, it follows that the iterates $x^{(k)}$ are each a function of the starting point $x^{(0)}$.

In limit of infinitely small step size, $\epsilon \to 0$, the local descent iterates $x^{(1)}(x^{(0)}), x^{(2)}(x^{(0)}), \ldots$, trace out a continuous trajectory $x(x^{(0)}, t)$, defined by

$$x_t(x_0) = \lim_{\epsilon \to 0} x^{(\lfloor t/\epsilon \rfloor)}(x_0).$$

The *length* of the trajectory is defined as

$$S(x_0) = \int_0^\infty ||\nabla x_t(x_0)|| dt.$$

Therefore, a sufficient condition for *non-convergence* to a global minimum x^* is that

$$S(x_0) \geq ||x_0 - x^*||.$$

Hence, one can get a sense of the difficulty of the problem by estimating the trajectory lengths. Intuitively, if most trajectory lengths are small, this indicates that the problem is riddled with local minima.

Suppose we initialize the local descent at a random point $x^{(0)}$, where the law of $x^{(0)}$ has density $\mu^0(x)$ with respect to Lesbegue measure. Then the marginal density $\mu^t(x)$ of the current iterate at time t is given by the *continuity equation*:

$$\frac{\partial}{\partial t} \mu^t(x) = -\nabla \cdot (f'(x) \mu^t(x)).$$

Considering the densities μ^t allows one to relate local properties, such as trajectory length, to statistics of the local minima. For instance, one can compute the *average path length* via the integral

$$\mathbf{E}_{\mu^0}[S(x_0)] = \int_0^\infty \int_{\mathcal{X}} ||g(x)|| \mu^t(x) dx dt.$$

Question: Can we obtain bounds for the average path length, if, for example, f is a Gaussian process?

2 Example: Sparse recovery in the white-noise limit

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.

Consider a sparse recovery problem with parameters $\theta \in \mathbb{R}^d$ corresponding to locations of point sources, where the signal a source at θ is given by

$$\psi_\theta(x) = \psi(x - \theta),$$

where $\psi : \mathbb{R}^p \to \mathbb{R}^+$ is a nonnegative kernel function.

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

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

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 $\psi_\theta(x)$ 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

$$y_i = m(x_i) + \epsilon_i,$$

where ϵ_i are indedependent with mean 0 and variance σ^2.

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

$$\text{minimize} \sum_{i=1}^{n} \left\| y_i - \sum_{j=1}^{K} w_j \psi_{\theta_j}(x_i) \right\|^2 .$$

3 References

- Ambrosio, Luigi, Nicola Gigli, and Giuseppe Savar. Gradient flows: in metric spaces and in the space of probability measures. Springer Science & Business Media, 2008.

- Morgenshtern, Veniamin I., and Emmanuel J. Candes. "Super-Resolution of Positive Sources: the Discrete Setup." arXiv preprint arXiv:1504.00717 (2015).