Reporte: Convex Optimization for Big Data

Victor Quintero Marmol Gonzalez 175897

31 de mayo de 2018

Abstracto

En este artículo escrito por *Volkan Cevher*, *Stephen Becker* y *Mark Schmidt* en el 2014, se revisa los recientes avances en algoritmos de optimización convexa para grandes cantidades de datos (Big Data), los cuales tienen como intención reducir los cuellos de botella computacionales, de almacenamiento y comunicación. Se habla de técnicas contemporaneas como métodos de primer orden y aleatorización para la escalabilidad, y se estudia la importancia que tiene la computación paralela y distribuida. Estos algoritmos se basan en principios simples y logran asambrar por sus resultados aun en problemas clásicos.

Optimización convexa en el surgimiendo de Big Data

Los autores mencionan que la importancia de las formulaciones convexas y la optimización ha aumentado dramáticamente en la última década. Describen que hay varias razones importantes para esta explosión de interés, dos de las más obvias son la existencia de algoritmos eficientes para el cálculo de soluciones globalmente óptimas y la capacidad de usar geometría convexa para probar propiedades útiles sobre la solución

Sin embargo también se habla que esta nueva popularidad en la optimización convexa coloca los algoritmos convexos en una situación de presión para acomodar conjuntos de datos cada vez más grandes y para resolver problemas en dimensiones sin precedentes. Problemas de como información de internet, de texto o de imágenes pasaron de producir datos de gigabytes a terabytes o incluso exabytes.

Se hace mención que la optimización convexa se esta reinventando en respuesta para tratar estos problemas de grandes cantidades de datos y parámetros, lo que los hace muy dificil de procesar localmente. Incluso rutinas básicas de álgebra lineal como descomposiciones de Cholesky y multiplicaciones matriz-matriz o matriz-vector que toman los algoritmos otorgados son prohibitivos. Otro punto importante mencionado es que los algoritmos convexos tampoco necesitan buscar soluciones de alta precisión ya que los modelos de Big Data son por definición simples o inexactos.

Bases

Los autores describen los fundamentos de la optimización Big Data a través de la siguiente formulación compuesta:

$$F* = min_x\{F(x) = f(x) + g(x) : x \in R^p\}$$

Donde f y g son funciones convexas.

Se hace mención que una comprensión básica de los algoritmos de optimización para Big Data como el anterior se basa en tres pilares clave:

- Métodos de primer orden
- Aleatorización
- Computación paralela y distribuida

Se hace mención de igual manera que estos tres conceptos se complementan para ofrecer sorprendentes beneficios de escalabilidad para la optimización de grandes bases de datos.

Métodos de primer orden

Los autores mencionan que estos métodos obtienen soluciones numéricas de precisión baja o media. Estos métodos cuentan con casi dimensiones independientes las tasas de convergencia, son teóricamente robustas a las aproximaciones de sus oráculos, y típicamente se basan en primitivas computacionales que son ideales para computación distribuida y paralela.

Una fuente principal de problemas de Big Data, y que motivó a los autores a realizar este artículo, es el modelo de observación lineal omnipresente en varias disciplinas:

$$y = \Phi x_0 + z$$

donde x_0 es un parámetro desconocido, $\Phi \in R^{n \times p}$ es una matriz conocida y $z \in R^n$ representa perturbaciones desconocidas o ruido. De esta manera se revisa en el artículo métodos de primer orden para la optimización convexa suave y para la optimización convexa no uniforme.

Técnicas de aleatorización

En palabras de los autores, las técnicas de aleatorización se destacan particularmente entre muchas otras técnicas de aproximación para mejorar la escalabilidad de los métodos de primer orden ya que se puede controlar su comportamiento esperado. Las ideas clave incluyen actualizaciones parciales aleatorias de variables de optimización, reemplazando el gradiente determinista y los cálculos proximales con estimadores estadísticos baratos, y acelerando las rutinas básicas de álgebra lineal mediante la aleatorización.

Esta sección se describen aproximaciones aleatorias emergentes que aumentan el alcance de los métodos de primer orden a escalas extraordinarias. Los autores se concentran solo en las funciones F suaves y fuertemente convexas. Algunos problemas notables de Big Data realmente satisfacen esta suposición, y se nos da como ejemplo el problema del PageRank de Google, que mide la importancia de los nodos en un gráfico dado a través de su matriz de incidencia.

Computación paralela y distribuida

Se menciona que los métodos de primer orden proporcionan una base flexible para distribuir tareas de optimización y realizar cálculos en paralelo. Sin embargo, se puede mejorar aún más estos métodos con aproximaciones para aumentar los niveles de escalabilidad, desde algoritmos paralelos sincrónicos idealizados con comunicaciones centralizadas a algoritmos asíncronos enormemente escalables con comunicaciones descentralizadas.

Además en esta sección se describe varios desarrollos clave relacionados con los métodos de primer orden dentro del contexto de comunicación y sincronización. Se hace mención además de que se dajó fuera de este artículo problemas importantes que afectan el rendimiento práctico de estos métodos, como la latencia y los esquemas de comunicación de saltos múltiples, debido a falta de espacio.

Conclusión

Este artículo, a pesar de ser un poco técnico, me sorprendio gratamente ya que refleja a lo que nos enfrentamos en nuestra vida laboral. Al salir de la licenciatura (y estoy seguro que me va a pasar saliendo de la maestría) nos enfrentamos a que la mayoría de los problemas del mundo real son mucho más complicados y con muchas más restricciones que los vistos en clases, por lo que siempre es bueno estar al día y conocer nuevas metodologías e investigaciones que estan desarrollando diferentes personas sobre esta nueva ola de grandes bases de datos.