Convex Optimization for Big Data: Resumen

Héctor Adolfo Corro Zárate[1]

Maestría en Ciencia de Datos
Instituto Tecnológico Autónomo de México

1. Introducción

Este artículo muestra los avances recientes en los algoritmos de convexidad aplicados a Big Data, estos con la finalidad de reducir el cálculo, uso de memoria y cuellos de botella en su ejecución. De igual manera se describe acerca de algunas aproximanciones o técnicas contemporáneas como métodos de primer orden, aleatorización para escalabilidad y estudios breves sobre la importancia computo distribuido y paralelo de los temas anteriores.

2. Convexidad en Big Data

Dado el incremento sobresaliente en teorías sobre estructuración espaciada y minimización de rangos, así como modelos de aprendizaje estadísiticos como las máquinas de soporte vectorial, la importancia de formulaciones convexas y de optimización ha crecido en la última década. Entre varias de las razones de lo anterior se encuentran las siguientes dos cuestiones: la existencia de algoritmos eficientes para el computo global de soluciones optimas y la habilidad o característica de usar la geometria convexa para probar lo util de las propiedas intrínsecas de la solución.

Contrario a lo que sucedia en el pasado, tener que analizar grandes conjuntos de datos ya sean todos o en su mayoría para encontrar la solución más precisa y optima, los algoritmos convexos no necesitan ya encontrar esa alta precisión en sus soliuciones desde que los modelos de Big Data son simples e inexactos.

3. Lo básico

En esta parte del artículo se explica o muestra el entendimiento básico de los algoritmos de optimización para Big Data y en que conceptos recaen estos:

- **Métodos de primer orden**: aquellos que logran una precisión media en su solución numérica mediante el uso único de oraćulos de primer orden de información desde el objetivo tales como los gradientes de estimación.
- **Aleatorización**: las técnicas de aleatorización particularmente destacan de entre muchas otras técnicas de aproximación para mejorar la escalabilidad de los métodos de primer orden gracias a que se puede controlar su compartamiento esperadod de estos.
- **Computo paralelo y distribuido**: aquí los métodos de primer orden proveen casi de manera natural un esquema flexible para distribuir las tareas de optimización y ejecutar un computo en paralelo.

Los conceptos que se presentan anteriormente se complementan entre si para ofrecer escalabilidad en la optimización del Big Data.

4. Ejemplos

En una amplia segunda sección del artículo se mencionan y ejemplifican métodos y técnicas desprendidas de los 3 puntos anteriores que se ven a continuación:

Métodos de primer orden: en esta sección se menciona o repasa el operador LASSO que se presenta a continuación:

$$\hat{x}_{LASSO} = argmin\{F(x) := \frac{1}{2}\|y - \Phi x\|_2\} \tag{1}$$

Este operador muesta la ventaja, según el artículo, de proveer soluciones dispersas.

En una tabla ,que igual se muestra, se ven las diferencias de los operadores como LASSO y aquellos clásicos como puntos interiores; en los que se ve un mejor desempeño como en la casi dimensionalidad-independencia, convergencia y la ventaja de explotar operadores lineales implícitos. En contraste con los métodos de puntos interiores que requieren mucho espacio y poseer dimensionalidad dependiente a pesar de podersele aplicar diferentes operaciones matriciales previamente.

- **Objetivos de suavización**: Otro punto abordado dentro de los métodos de primer orden son aquellos objetivos de suavización. Son aquellos en los que la funcioón objetivo F solo consiste de una función covexa diferenciable f. Una técnica elemental para este tipo de casos cpnsiste en el método del gradiente que usa el gradiente local $\nabla f(x)$ y que iterativamente calcula lo siguiente:

$$x^{k+1} = x^k - \alpha \nabla f(x^k) \tag{2}$$

Para una minimización suave se recomienda usar otros algoritmos más rápidosçomo aquellos "Newton". Sin embargo, estos no se abordaron por carecer de información sobre F.

Al hacer simples suposiciones sobre f de manera rigurosa se podría analizar cuantas iteraciones necesitará el gradiente para llegar a la solución optima. Una suposición común que se mantiene es que el gradiente de f es Lipschitz continuo, lo que siginfica:

$$\forall x, y \in \mathbb{R}^p, \|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2 \tag{3}$$

Para una constante L. Cuando f es doblemente diferenciabl una condición suficiente son los eigenvalores de su hessiano $\nabla^2 f(x)$ acotado por arriba por L. Por lo tanto se podría estimar que $L = \|\Phi\|_2$. Y simplemente se establece al ciclo como $\alpha_k = \frac{1}{L}$.

Nesterov mencionará algo de mejorar lo anterior y proponer otra técnica.

- **Objetivos Compuestos**: La función objetivo F consiste de una función convexa y diferenciable f y una función no-suavizada conexa g. En general se menciona que g parece reducir sustancialmente la eficiencia de los métodos de primer orden aquí expuestos.

  En la siguiente figura se verá que no se puede estar más equivocado.

- **Objetivos proximos**: al llegar a esta parte del artículo el autor menciona que los anteriores vistos no son "tan"directamente aplicables y se proponen entonces este tipo de técnicas.

$$min_{x,z \in \mathbb{R}^p}\{F(x.z) = h(x) + g(z) : \Phi z = x\} \tag{4}$$

  De ver la ecuación anterior se desprende que se pueden mejorar tanto los modelos como la capacidad de computo de los procesos. Es en este punto cuando se propone el algoritmo ADMM que apalanca con poderosos lagrangeanos y tecnicas de descomposición dual.

Escalabilidad via aleatorización de Big Data: En esta parte del artículo se menciona y da reconocimiento a los métodos de primer orden como solución a problemas de gran escala. Sin embargo, en la práctica el cálculo numérico exacto de estas soluciones demanda de muchas iteraciones lo que puede provocar que sean inviables y que la dimensión de estos crezca.

En esta sección se describen varias aproximaciones de aleatorización que incrementan el alcance de los métodos de primer orden.

- **Métodos descendentes coordinados**: en esta parte se hace hincapie en seleccionar de manera adecuada de cada i iteración sobre la ecuación 2. Y pide observar los siguientes pasos:

  1. Escoger un índice $i_k \in \{1, 2, ..., p\}$
  2. $x^{k+1} = x^k - \alpha \nabla_{ik} F(x^k) e_{ik}$

  Lo anterior ejemplifica la dificultad el los métodos descendientes coordinados. Encontrar la mejor coordinación para actualizar el máximo del gradiente de las magnitudes de cada elemento puede requerir de mucha capacidad de computo. Una alternativa de lo anterior es iterar a través de todos los coordinamientos de maner secuencial. Esto implica la estrategia más "barata.[en] cuanto a resultados de indices de convergencia.

- **Métodos del Gradidente Estocástico (Stochastic Gradient Methods**: en contraste con el punto anterior en el que los métodos atualizan un coordinado a la vez con su gradiente exacto, los métodos estocásticos actualizan todos los coordinados simultaneamente pero usan gradientes aproximados. Al igual que los métodos descendentes coordinados el diseño crucial del problema para el gradiente estocástico consiste en la correcta selección de los puntos de datos j en cada iteración.

- **Aleatorización del Álgebra Lineal (Randomized linear algebra)**: al tratarse de operaciones matriciales como descomposición de estas, multiplicación de matrices esto puede resultar en cuellos de botella y mayor computo. La idea propuesta de esto es acercar los métodos lo más que se pueda a la forma $M \approx Q(Q^T M)$ con $Q \in \mathbb{R}^{pxr}$ o construir una representación de rango bajo ya se por selección de filas o columnas en orden de acelerar el cálculo. De hecho haciendo lo anterior obtiene el control sobre la distribución de los errores. Como segunda parte a este punto propone un algoritmo de aproximación de rango bajo como una posible propuesta de solución, con menores iteraciones y con métodos de primer orden exactos mientras conserva precisión en el objetivo.

5. El rol del computo paralelo y distribuido

El autor indica que mientras los métodos de primer orden parecen ser ideales para desempeños casi optimos, 2 asuntos parecen interferir cuando se hace uso de hardware heterogeneo y distribuido.

- **Comunicación**
- **Sincronización**

Conforme se sigue leyendo acerca de este tema, los métodos de primer orden se pueden beneficiar ampliamente del computo paralelo. Estos sistemas de computo se encuentran tipificados por nodos de procesamiento uniformes que se encuentran próximos y tienen una alta fiabilidad en cuanto a comunicación.

Métodos de primer orden con comunicación reducida o comunicaciones decentralizadas: la ventaja de esto, menciona el autor, en términos de comunicación es que cada procesador solo necesita comunicarse con un solo punto de coordinación mientras que solo necesita recibir de las actualizaciones de los coordinados que han cambiado.

6. Lo que viene

Como resumen de lo que viene, el autor puntualiza que para resolver problemas de optimización convexa con recursos computacionales modestos es necesario identificar la estructura clave de la dependencia de los algoritmos y sus elementos escenciales para un costo-beneficio adecuado. Al ser la sincronización y comunicación de los procesos restricciones del hardware disponible en la actualidad esto determina la naturaleza de los algoritmos elegidos.