**Convex optimization for big data**

(Cevher, Becker y Schmidt)

2014

Luis Federico Puente Peña

103108

**Objetivo**

El objetivo del documento es revisar los avances recientes para desarrollar algoritmos de optimización convexos para *big data*; los cuales buscan reducir los cuellos de botella computacionales, de almacenamiento y de comunicaciones.

El documento menciona que los algoritmos básicos de optimización, utilizando *big data*, se basan en tres pilares: métodos de primer orden, escalar vía aleatorización y cómputo en paralelo y distribuidos. Los cuales se describen a continuación.

1. **Métodos de primer orden**

Los métodos de primer orden utilizan las soluciones numéricas a partir de los métodos de optimización, por ejemplo, utilizando el gradiente. Funcionan incluso para funciones que no son suaves mediante el análisis de vecindad (*proximal mapping*).

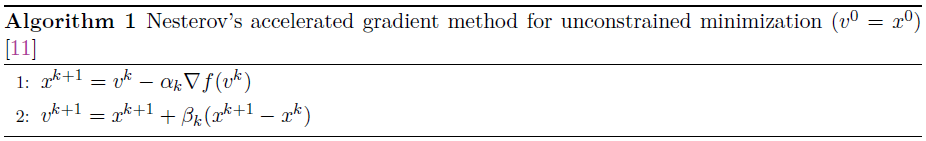
En estos métodos es posible incluir regularizaciones, incluso a funciones no suaves y pueden desempeñar un papel indispensable en la calidad de la solución. Considerando el método de “descenso por gradiente” puede verse que problemas no suaves pueden resolverse casi tan eficientemente como sus contrapartes suaves.

**Objetivos suaves**

El método de gradiente consiste en calcular el gradiente en un punto (local) y avanzar una determinada unidad (tamaño de paso) en la dirección de mayor crecimiento/decrecimiento y posteriormente realizar iteraciones.

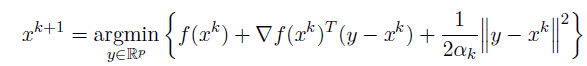


El método de gradiente acelerado logra la mejor tasa de error posible del peor de los casos, y por lo tanto, se lo conoce como un método óptimo de primer orden.



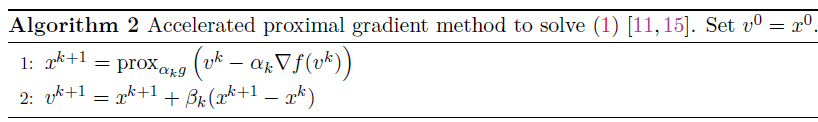
**Objetivos compuestos**

En esta sección se analiza el problema compuesto canónico, donde la función objetivo consiste en una función convexa diferenciable f y una función convexa no suave g. Se basa en una aproximación cuadrática local simple de f:



Los objetivos compuestos están lejos de ser problemas genéricos de optimización convexa no suave. Los métodos de descenso por gradiente aprovechan la estructura compuesta para mantener las mismas tasas de convergencia del método de gradiente para las clases de problemas suaves.

El método acelerado de gradiente proximal se define de forma análoga:



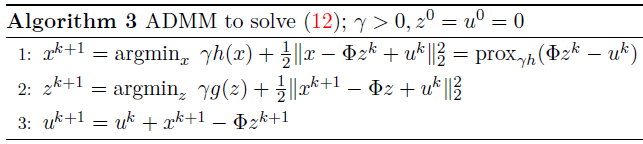
**Objetivos de proximidad**

Para muchas aplicaciones, los métodos de primer orden hora no son directamente aplicables. Como resultado, encontraremos útil considerar la siguiente forma funcional:

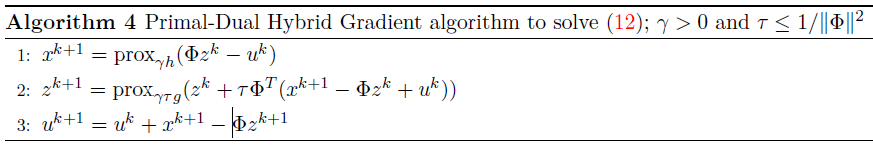


Estableciendo que los operadores de proximidad de h y g son ambos eficientes.

Es posible aplicar un algoritmo simple, denominado método de multiplicación de multiplicadores (ADMM) para sus soluciones, que aprovecha poderosas técnicas aumentadas de Lagrange y descomposición dual.



Hay dos inconvenientes. Hay ocasiones en que la matriz no es diagonalizable y puede ocurrir que no exista convergencia. Para solucionar esto se puede plantear el siguiente algoritmo:



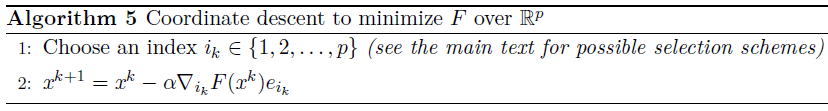
1. **Escalar vía aleatorización**

Las técnicas de aleatorización mejoran la escalabilidad de los métodos de primer orden ya que controlan su comportamiento esperado. Incluyen actualizaciones parciales aleatorias de variables de optimización, reemplazando el gradiente determinista y los cálculos proximales con estimadores estadísticos y acelerando las rutinas básicas de álgebra lineal mediante la aleatorización.

En la práctica, sin embargo, los cálculos numéricos y las iteraciones de los métodos de primer orden pueden hacer que incluso estos métodos simples sean inviables a medida que crecen las dimensiones del problema. A continuación se describe las aproximaciones aleatorias emergentes que aumentar el alcance de los métodos de primer orden.

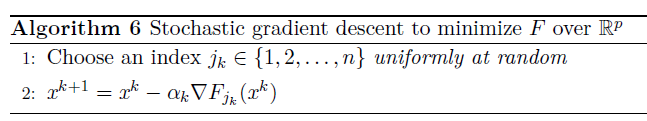
**Coordinar los métodos de descenso**

El cálculo del gradiente completo para la formulación del problema del PageRank requiere una operación matriz-vector en cada iteración. Una operación de vector más económica sería elegir una coordenada i de x y solo modificar la variable correspondiente xi para mejorar la función objetivo. Esta idea captura la esencia de los métodos de descendencia coordinada, que tienen una larga historia en la optimización.



**Métodos de gradiente estocástico**

En contraste con los métodos de descendencia de coordenadas aleatorizados, que actualizan una sola coordenada a la vez con su gradiente exacto, los métodos de gradiente estocástico actualizan todas las coordenadas simultáneamente pero usan gradientes aproximados.

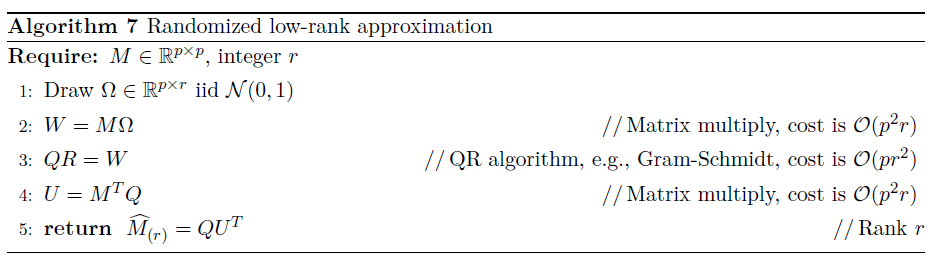


De forma similar a los métodos de descendencia coordinada, el problema de diseño crucial en los métodos de gradiente estocásticos es la selección de los puntos de datos j en cada iteración.

**Álgebra lineal aleatorizada**

Para problemas de Big Data, las operaciones de álgebra lineal básica, como las descomposiciones de matrices (por ejemplo, valor propio, valor singular y Cholesky) y las multiplicaciones matriciales pueden ser importantes cuellos de botella computacionales debido a su dependencia lineal de las dimensiones. Sin embargo, cuando los objetos de matriz relevantes tienen representaciones de bajo rango, la eficiencia de estos métodos mejora de manera uniforme.

La idea detrás de los métodos aleatorizados de álgebra lineal es aproximar M= Q (QTM) con Q en , o para construir una representación de bajo rango por selección de subconjunto de columna o fila para acelerar el cálculo.

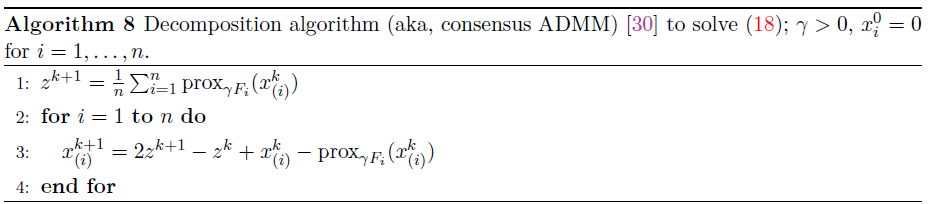


1. **Cómputo en paralelo y distribuidos**

Consisten en algoritmos paralelos sincrónicos idealizados con comunicaciones centralizadas a algoritmos asíncronos enormemente escalables con comunicaciones descentralizadas.

**Métodos de primer orden violentamente paralelos**

Los métodos de primer orden pueden beneficiarse significativamente de la computación en paralelo. La expresión violentamente paralela se refiere a un escenario ideal para la paralelización en el que dividimos el trabajo en cálculos independientes que se pueden realizar simultáneamente de manera predecible. Cada procesador se comunica con la ubicación central para formar el gradiente final y lograr la velocidad lineal ideal.



**Métodos de primer orden con comunicaciones reducidas o descentralizadas**

En sistemas grandes, comunicar el gradiente o sus elementos a una ubicación central puede crear un cuello de botella en la comunicación. En esta configuración, los métodos de descendencia coordinada proporcionan un enfoque basado en principios para reducir las comunicaciones. La idea básica es aplicar varias actualizaciones de descenso coordinado al mismo tiempo en paralelo.

**Métodos asíncronos de primer orden con comunicaciones descentralizadas**

El gradiente y los métodos de descomposición anteriores aún requieren una sincronización global para manejar problemas separables

Bajo ciertas condiciones, este procedimiento asíncrono preserva la convergencia de los métodos de gradiente estocástico y da como resultado una aceleración sustancial cuando hay muchos núcleos disponibles. El mismo modelo sin bloqueo de memoria también se aplica a los métodos de descenso de coordenadas paralelas estocásticas. Además, los algoritmos de primer orden con aleatorización pueden ser efectivos incluso en configuraciones asíncronas y descentralizadas con la posibilidad de fallas de comunicación.

1. **Consideraciones**

Los problemas de Big Data requieren una revisión fundamental de cómo diseñamos algoritmos de optimización convexos y sugieren opciones computacionales no convencionales.

Dado que las restricciones de sincronización y comunicación del hardware disponible naturalmente dictan la elección de los algoritmos, se espera que sigan descubriendo nuevas herramientas de aproximación que idealmente adapten los algoritmos convexos a la heterogeneidad de las plataformas computacionales. También se predice una mayor utilización de modelos compuestos y los correspondientes principios de mapeo proximal para problemas de *big data* no suaves para hacer frente al ruido y otras limitaciones.