Modern Optimization Methods for Big Data Problems

MATH11146
The University of Edinburgh
Spring 2017

Peter Richtárik

Modern Optimization Methods for Big Data Problems

Lecture 5

Stochastic Dual Subspace Ascent

February 6, 2017

Motivation

- Recall that assuming exactness, and under certain assumptions in the stepsize ω, the iterates of the **basic method** converge[4] in the weak sense (Theorem 29) and/or in the strong sense (Theorem 36) to

$$x_* \overset{\text{def}}{=} \Pi^{\mathbf{B}}_{\mathcal{L}}(x_0).$$

- That is, the basic method actually solves the optimization problem:

$$\begin{array}{ll} \text{minimize} & P(x) \overset{\text{def}}{=} \frac{1}{2}\|x - x_0\|^2_{\mathbf{B}} \\ \text{subject to} & \mathbf{A}x = b \\ & x \in \mathbb{R}^n. \end{array} \tag{66}$$

  We will call (66) the **primal problem**, and P the **primal objctive function.**

- In optimization, one can associate with each optimization problem a closely related optimization problem, called the **dual problem.**
- We shall now investigate several very interesting relationships between the primal and the dual problems.

[4]This is also true for the parallel and accelerated methods. However, we shall not deal with them in this lecture.

Duality

Dual Problem: Concave Quadratic Maximization

The **dual problem** to (66) is the optimization problem

$$\begin{array}{ll} \text{maximize} & D(y) \overset{\text{def}}{=} (b - \mathbf{A}x_0)^\top y - \frac{1}{2}\|\mathbf{A}^\top y\|^2_{\mathbf{B}^{-1}} \\ \text{subject to} & y \in \mathbb{R}^m. \end{array} \tag{67}$$

- $D : \mathbb{R}^m \to \mathbb{R}$ is the **dual objective function** (quadratic)
- The dimension of the dual variable (y) is m (# rows of $\mathbf{A}$).
  The dimension of the primal variable (x) is n (# columns of $\mathbf{A}$).
- A more detailed look at the terms:
  - The first term, $(b - \mathbf{A}x_0)^\top y$, is linear in y.
  - The second term can be written as $-\frac{1}{2}y^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y$.
  - Thus, the **gradient and Hessian of D** are given by:

$$\nabla D(y) = b - \mathbf{A}x_0 - \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y, \qquad \nabla^2 D(y) = -\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \tag{68}$$

- Note that $\nabla^2 D(y)$ is a **negative semidefinite matrix.** Equivalently, $-\nabla^2 D(y)$ is a **positive semidefinite matrix.** Hence:
  - D is a concave quadratic function
  - $-D$ is a convex quadratic function

Weak Duality

Lemma 40 (Weak Duality)

For any **primal feasible point** x *(i.e., $x \in \mathbb{R}^n$ for which $\mathbf{A}x = b$) and for any* **dual feasible point** *(i.e., $y \in \mathbb{R}^m$), we have*

$$P(x) \geq D(y).$$

Proof.

For any $x \in \mathbb{R}^n$ for which $\mathbf{A}x = b$ and for any $y \in \mathbb{R}^m$ we have

$$\begin{aligned} P(x) - D(y) &\overset{(66)+(67)}{=} \tfrac{1}{2}\|x - x_0\|^2_{\mathbf{B}} + \tfrac{1}{2}\|\mathbf{A}^\top y\|^2_{\mathbf{B}^{-1}} + (x_0 - x)^\top \mathbf{A}^\top y \\ &= \tfrac{1}{2}\|\mathbf{B}^{1/2}(x - x_0)\|^2 + \tfrac{1}{2}\|\mathbf{B}^{-1/2}\mathbf{A}^\top y\|^2 + (x_0 - x)^\top \mathbf{A}^\top y \\ &= \tfrac{1}{2}\|\mathbf{B}^{-1/2}\mathbf{A}^\top y + \mathbf{B}^{1/2}(x_0 - x)\|^2 \\ &= \tfrac{1}{2}\|x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y - x\|^2_{\mathbf{B}} \geq 0. \end{aligned}$$

$\square$

Optimality Conditions

Definition 41 (Duality Mapping)

The **duality mapping** is the function $x(y) : \mathbb{R}^m \to \mathbb{R}^n$ defined by

$$x(y) \overset{\text{def}}{=} x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y. \tag{69}$$

Theorem 42

(i) **Dual boundedness.** *D is bounded above $\Leftrightarrow$ the primal problem is feasible*

(ii) **Dual optimality.**

$$y \text{ is dual optimal} \quad \Leftrightarrow \quad \mathbf{A}x(y) = b \tag{70}$$

(iii) **Primal optimality.**

$$x = x_* \quad \Leftrightarrow \quad \mathbf{A}x = b \quad \text{and} \quad x = x(y) \text{ for some } y \tag{71}$$

(iv) **x_* can be obtained from any dual optimal point:**

$$y_* \text{ is dual optimal} \quad \Rightarrow \quad x(y_*) = x_* \tag{72}$$

Proof of Theorem 42

(i) Since D is a concave quadratic function, it has a maximizer if and only if there exists y such that $\nabla D(y) = 0$ (in which case any such y is a maximizer). In view of (68), this happens if and only if the following linear system has a solution:

$$\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y = b - \mathbf{A}x_0. \tag{73}$$

This system has a solution if and only if

$$b - \mathbf{A}x_0 \in \mathrm{Range}\left(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top\right) \overset{\text{Fact 17(iii)}}{=} \mathrm{Range}\left(\mathbf{A}\right).$$

Finally, this happens if and only if $b \in \mathrm{Range}(\mathbf{A})$, which means that the primal problem is feasible.

(ii) Using the reasoning in (i), we know that y is dual optimal $\Leftrightarrow$ y solves (73). It remains to notice that (73) can equivalently be written as $\mathbf{A}x(y) = b$.

(iii) Do this as an exercise. *Hint:* Use weak duality; in particular, the derived expression for $P(x) - D(y)$.

(iv) This follows by combining (ii) and (iii).

Dual Suboptimality vs Primal Suboptimality

The dual-to-primal mapping enjoys the following insightful property:

Theorem 43

Let y_* be any dual optimal point and $y \in \mathbb{R}^m$. Then

$$D(y_*) - D(y) = \tfrac{1}{2}\|x_* - x(y)\|^2_{\mathbf{B}}. \tag{74}$$

Proof.

$$\begin{aligned}
D(y_*) - D(y) &\overset{(67)}{=} (b - \mathbf{A}x_0)^\top (y_* - y) - \tfrac{1}{2} y_*^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y_* + \tfrac{1}{2} y^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y \\
&\overset{(70)}{=} y_*^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top (y_* - y) - \tfrac{1}{2}(y_*)^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y_* \\
&\quad + \tfrac{1}{2} y^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top y \\
&= \tfrac{1}{2}(y - y_*)^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top (y - y_*) \\
&\overset{(69)}{=} \tfrac{1}{2}\|x(y) - x(y_*)\|^2_{\mathbf{B}}.
\end{aligned}$$

It remains to use (72), which states that $x(y_*) = x_*$. □

Dual Algorithms Solve the Primal Problem

Let $\{y_k\}_0^\infty$ be any sequence for which

$$D(y_k) \to D(y_*).$$

Such a sequence can be obtained by **any algorithm that solves the dual problem.** In view of Theorem 43, we automatically have

$$x(y_k) \to x(y_*) = x_*.$$

Now, define an associated **primal algorithm** via the iterates:

$$x_k \overset{\text{def}}{=} x(y_k). \tag{75}$$

Conclusion: **any convergent dual algorithm automatically leads to a convergent primal algorithm.**

Stochastic Dual Subspace Ascent

Algorithm: Stochastic Dual Subspace Ascent (SDSA)

Consider the following algorithm for solving the dual problem (67):

$$\boxed{y_{k+1} = y_k + \mathbf{S}_k \lambda_k} \tag{76}$$

$\mathbf{S}_k$ is a fresh sample from $\mathcal{D}$, and λ_k is a suitably chosen **"stepsize" parameter.** We refer to this method by the name **stochastic dual subspace ascent (SDSA).**

- **Why stochastic?** Because the iterates are random vectors, which follows from the fact that $\mathbf{S}_k$ is a random matrix.
- **Why subspace?** The step, $\mathbf{S}_k \lambda_k$, can potentially be any point in a specific **random subspace** of $\mathbb{R}^m$. In particular, this is the space $\mathrm{Range}(\mathbf{S}_k)$, i.e., the subspace spanned by the columns of the random matrix $\mathbf{S}_k$. We hope that by focusing on a random subspace (of a sufficiently small dimension) in each iteration, we can perform the iteration much faster, particularly if m is big.
- **Why ascent?** We wish the method to always improve the dual function value (or, at least, not to make it worse): $D(y_{k+1}) \geq D(y_k)$. We achieve this by an appropriate choice of λ_k. In particular, in SDSA we pick the best vector λ_k; i.e., the vector for which $D(y_k + \mathbf{S}_k \lambda_k)$ is maximized!

How to Compute the Best λ_k? I

In SDSA we pick the stepsize parameter λ_k via

$$\lambda_k \overset{\text{def}}{=} \operatorname{argmax}_\lambda D(y_k + \mathbf{S}_k\lambda).$$

Since the function $\psi(\lambda) = D(y_k + \mathbf{S}_k\lambda)$ is a concave quadratic, λ is its maximizer if and only if

$$\nabla\psi(\lambda) = 0. \tag{77}$$

Since

$$\begin{aligned}
\nabla\psi(\lambda) &= \mathbf{S}_k^\top \nabla D(y_k + \mathbf{S}_k\lambda) \overset{(68)}{=} \mathbf{S}_k^\top \left(b - \mathbf{A}x_0 - \mathbf{AB}^{-1}\mathbf{A}^\top(y_k + \mathbf{S}_k\lambda)\right) \\
&= \mathbf{S}_k^\top \left[b - \mathbf{A} \underbrace{\left(x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_k\right)}_{\overset{(69)}{=} x(y_k)} \right] - \mathbf{S}_k^\top \mathbf{AB}^{-1}\mathbf{A}^\top \mathbf{S}_k\lambda \\
&= \mathbf{S}_k^\top \left(b - \mathbf{A}x(y_k)\right) - \mathbf{S}_k^\top \mathbf{AB}^{-1}\mathbf{A}^\top \mathbf{S}_k\lambda,
\end{aligned}$$

How to Compute the Best λ_k? II

equation (77) is equivalent to the **linear system:**

$$\mathbf{S}_k^\top \mathbf{AB}^{-1}\mathbf{A}^\top \mathbf{S}_k\lambda = \mathbf{S}_k^\top \left(b - \mathbf{A}x(y_k)\right). \tag{78}$$

If we wish to be greedy, we may choose λ_k as any solution of the linear system (78). In SDSA, we pick a **particular solution** of (78): **the least-norm solution.** In view of Exercise 5, the least-norm solution of a linear system is given by applying the pseudoinverse of the system matrix to the right hand side. Thus, we get:

$$\begin{aligned}
\lambda_k &\overset{\text{def}}{=} \arg\min_\lambda \{\|\lambda\| \;:\; (78)\ \textit{holds}\} \\
&\overset{\text{Exercise 5}}{=} \left(\mathbf{S}_k^\top \mathbf{AB}^{-1}\mathbf{A}^\top \mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top \left(b - \mathbf{A}x(y_k)\right).
\end{aligned} \tag{79}$$

Plugging this back into the SDSA iteration, we get

$$\boxed{y_{k+1} \overset{(76)+(79)}{=} y_k - \mathbf{S}_k \left(\mathbf{S}_k^\top \mathbf{AB}^{-1}\mathbf{A}^\top \mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top \left(\mathbf{A}x(y_k) - b\right)} \tag{80}$$

Duality of SDSA and the Basic Method with Unit Stepsize

A natural question: How do the iterates of the primal algorithm (defined in (75)) associated with the dual iterates of SDSA (defined in (80)) look like?

$$
\begin{aligned}
x(y_{k+1}) &\overset{(69)}{=} x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_{k+1} \\
&\overset{(80)}{=} \underbrace{x_0 + \mathbf{B}^{-1}\mathbf{A}^\top y_k}_{\overset{(75)}{=} x(y_k)} - \mathbf{B}^{-1}\mathbf{A}^\top \underbrace{\mathbf{S}_k \left(\mathbf{S}_k^\top \mathbf{A}\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}_k\right)^\dagger \mathbf{S}_k^\top}_{\mathbf{H}_k} (\mathbf{A}x(y_k) - b) \\
&= x(y_k) - \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}_k(\mathbf{A}x(y_k) - b).
\end{aligned}
$$

Observe:

- If we set $y_0 = 0$, then $x(y_0) = x_0$
- **This is the basic method with unit stepsize!** (see (7))

Theorem 44 (The Basic Method with Unit Stepsize is a "Mirror Image" of SDSA)

Let $y_0 = 0$ and let $\{y_k\}$ be the iterates (80) of SDSA. Then the primal iterates $x_k = x(y_k)$ associated with SDSA exactly correspond to the basic method with unit stepsize ($\omega = 1$).

Convergence of SDSA

By applying Theorem 43 to SDSA (with starting point $y_0 = 0$) and iterates $\{y_k\}$, we get

$$D(y_*) - D(y_k) = \tfrac{1}{2}\|x_* - x_k\|_{\mathbf{B}}^2,$$

where in view of Theorem 44, $\{x_k\}$ are the iterates of the the basic method with unit stepsize.

By taking expectations on both sides of the above identity, we get

$$\mathrm{E}\left[D(y_*) - D(y_k)\right] = \tfrac{1}{2}\mathrm{E}\left[\|x_k - x_*\|_{\mathbf{B}}^2\right]. \tag{81}$$

By applying Theorem 36 (strong convergence of the basic method) to (81), with $\omega = 1$, we get:

Theorem 45 (Convergence of SDSA)

Choose any $x_0 \in \mathbb{R}^n$. Let Assumption 3 (exactness) hold and set $x_* = \Pi_{\mathcal{L}}^{\mathbf{B}}(x_0)$. Let $y_0 = 0$ and $\{y_k\}_{k=0}^\infty$ be the random iterates produced by SDSA (see (80)). Further, let $t_k \overset{def}{=} \mathrm{E}\left[D(y_*) - D(y_k)\right]$. Then for all $k \geq 0$ we have

$$(1 - \lambda_{\max})^k t_0 \leq t_k \leq (1 - \lambda_{\min}^+)^k t_0. \tag{82}$$

Special Cases: $\mathbf{S}_k$ is a Random Vector

If $\mathbf{S}_k$ has a single column only, then SDSA is moving in the **random direction $\mathbf{S}_k \in \mathbb{R}^m$,** using **stepsize $\lambda_k \in \mathbb{R}$.**

Special cases:

- If $\mathbf{S}_k$ is a **random coordinate vector,** i.e., if $\mathcal{D}$ is given by $\mathbf{S}_k = e_i$ (the ith unit basis vector in $\mathbb{R}^m$) with probability $p_i > 0$, then SDSA is called **stochastic dual coordinate ascent (SDCA).**
- If $\mathbf{S}_k$ is a **random Gaussian vector,** then SDSA is called **stochastic dual Gaussian ascent (SDGA).**