Modern Optimization Methods for Big Data Problems

MATH11146
The University of Edinburgh
Spring 2017

Peter Richtárik

Modern Optimization Methods for Big Data Problems

Lecture 3

Randomized Methods for Solving Linear Systems: Equivalence and Exactness

January 23, 2017

Projection and Pseudoinverse

Projection Operators and Pseudoinverse Matrices - I

Definition 7

The **B-pseudoinverse** of a matrix $\mathbf{M}$, is defined as

$$\mathbf{M}^{\dagger_{\mathbf{B}}} \stackrel{\text{def}}{=} \mathbf{B}^{-1}\mathbf{M}^{\top}(\mathbf{M}\mathbf{B}^{-1}\mathbf{M}^{\top})^{\dagger}, \tag{16}$$

where $\dagger$ denotes the standard pseudoinverse.

Exercise 4

Show that

(i) $\mathbf{A}^{\dagger}\mathbf{A}$ *is a symmetric matrix*

(ii) $\mathbf{A}^{\top}(\mathbf{A}\mathbf{A}^{\top})^{\dagger} = \mathbf{A}^{\dagger}$

(iii) *The* $\mathbf{I}$*-pseudoinverse is the standard Moore-Penrose pseudoinverse.*

Projection Operators and Pseudoinverse Matrices - II

Lemma 8

The projection onto $\mathcal{L} = \{x \ : \ \mathbf{A}x = b\}$ *is given by*

$$\Pi_{\mathcal{L}}^{\mathbf{B}}(x) = x - \mathbf{B}^{-1}\mathbf{A}^{\top}(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}^{\top})^{\dagger}(\mathbf{A}x - b) \overset{(16)}{=} x - \mathbf{A}^{\dagger_{\mathbf{B}}}(\mathbf{A}x - b). \quad (17)$$

Proof.

Do it yourself. □

Exercise 5

Show that $\mathbf{B}$*-pseudoinverse satisfies*

$$\mathbf{A}^{\dagger_{\mathbf{B}}} b = \Pi_{\mathcal{L}}^{\mathbf{B}}(0) = \arg\min_{x}\{\|x\|_{\mathbf{B}} \ : \ \mathbf{A}x = b\}.$$

Equivalence of Algorithms

Gradient and Hessian of $f_{\mathbf{S}}(x)$ - I

In order to keep the expressions as brief as possible throughout, it will be useful to define

$$\mathbf{Z} \overset{\text{def}}{=} \mathbf{A}^\top \mathbf{H} \mathbf{A} \overset{(13)}{=} \mathbf{A}^\top \mathbf{S} (\mathbf{S}^\top \mathbf{A} \mathbf{B}^{-1} \mathbf{A}^\top \mathbf{S})^\dagger \mathbf{S}^\top \mathbf{A}. \tag{18}$$

Lemma 9

$\mathbf{B}^{-1}\mathbf{Z}$ *is the projection, in the* $\mathbf{B}$*-norm, onto* $\mathrm{Range}\left(\mathbf{B}^{-1}\mathbf{A}^\top \mathbf{S}\right)$*. In particular,*

$$(\mathbf{B}^{-1}\mathbf{Z})^2 = \mathbf{B}^{-1}\mathbf{Z} \qquad \textit{and} \qquad \mathbf{Z}\mathbf{B}^{-1}\mathbf{Z} = \mathbf{Z}. \tag{19}$$

Recall from (3) that $f_{\mathbf{S}}(x) \overset{\text{def}}{=} \frac{1}{2}\|\mathbf{A}x - b\|^2_{\mathbf{H}} = \frac{1}{2}(\mathbf{A}x - b)^\top \mathbf{H}(\mathbf{A}x - b)$. By combining this with (18), this can be also written in the compact form

$$f_{\mathbf{S}}(x) = \frac{1}{2}(x - x_*)^\top \mathbf{Z}(x - x_*), \tag{20}$$

where x_* is any point in $\mathcal{L}$.

Gradient and Hessian of $f_{\mathbf{S}}(x)$ - II

Lemma 10

For each $x, h \in \mathbb{R}^n$ *we have the expansion*

$$f_{\mathbf{S}}(x + h) = f_{\mathbf{S}}(x) + \langle \nabla f_{\mathbf{S}}(x), h \rangle_{\mathbf{B}} + \tfrac{1}{2}\left\langle (\nabla^2 f_{\mathbf{S}})h, h \right\rangle_{\mathbf{B}},$$

where

$$\nabla f_{\mathbf{S}}(x) \overset{def}{=} \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b) \qquad \textit{and} \qquad \nabla^2 f_{\mathbf{S}} \overset{def}{=} \mathbf{B}^{-1}\mathbf{Z} \tag{21}$$

are the gradient and Hessian of $f_{\mathbf{S}}$ *with respect to the* $\mathbf{B}$*-inner product, respectively.*[2]

In view of (18) and (21), the gradient can also be written as

$$\nabla f_{\mathbf{S}}(x) = \mathbf{B}^{-1}\mathbf{Z}(x - x_*), \qquad x \in \mathbb{R}^n,\ x_* \in \mathcal{L}. \tag{22}$$

[2] If $\mathbf{B} = \mathbf{I}$, then $\langle \cdot, \cdot \rangle_{\mathbf{B}}$ is the standard Euclidean inner product, and we recover formulas for the standard gradient and Hessian. Note that $\mathbf{B}^{-1}\mathbf{Z}$ is both self-adjoint and positive semidefinite with respect to the $\mathbf{B}$-inner product. Indeed, for all $x, y \in \mathbb{R}^n$ we have $\left\langle \mathbf{B}^{-1}\mathbf{Z}x, y \right\rangle_{\mathbf{B}} = \langle \mathbf{Z}x, y \rangle_{\mathbf{I}} = \langle x, \mathbf{Z}y \rangle_{\mathbf{I}} = \left\langle x, \mathbf{B}^{-1}\mathbf{Z}y \right\rangle_{\mathbf{B}}$, and $\left\langle \mathbf{B}^{-1}\mathbf{Z}x, x \right\rangle_{\mathbf{B}} = \langle \mathbf{Z}x, x \rangle_{\mathbf{I}} \geq 0$.

Useful Identities Involving $f_{\mathbf{S}}(x)$

Lemma 11

For all $x \in \mathbb{R}^n$, we have

$$\begin{aligned} \nabla f_{\mathbf{S}}(x) &= (\nabla^2 f_{\mathbf{S}})\nabla f_{\mathbf{S}}(x) = (\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}}\nabla f_{\mathbf{S}}(x) \\ &= x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) = \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}(\mathbf{A}x - b). \end{aligned} \tag{23}$$

Moreover,

$$f_{\mathbf{S}}(x) = \frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}}. \tag{24}$$

Finally, if $\mathcal{L}_{\mathbf{S}}$ is the set of minimizers of $f_{\mathbf{S}}$, then $\mathcal{L} \subseteq \mathcal{L}_{\mathbf{S}}$, and

(i) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ f_{\mathbf{S}}(x) = 0\} = \{x \ : \ \nabla f_{\mathbf{S}}(x) = 0\}$

(ii) $\mathcal{L}_{\mathbf{S}} = x_* + \mathrm{Null}\left(\mathbf{B}^{-1}\mathbf{Z}\right)$ *for all* $x_* \in \mathcal{L}$

(iii) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}\mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top\mathbf{H}b\}$ *(see (4))*

(iv) $\mathcal{L}_{\mathbf{S}} = \{x \ : \ \mathbf{S}^\top\mathbf{A}x = \mathbf{S}^\top b\}$ *(see (6))*

Some Consequences of Lemma 11

- The identity $(\nabla^2 f_{\mathbf{S}})\nabla f_{\mathbf{S}}(x) = \nabla f_{\mathbf{S}}(x)$ means that the stochastic gradients of $f_{\mathbf{S}}$ are eigenvectors of the stochastic Hessian $\nabla^2 f_{\mathbf{S}}$, corresponding to eigenvalue one.
- The identity $(\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}}\nabla f_{\mathbf{S}}(x) = \nabla f_{\mathbf{S}}(x)$ means that the stochastic gradients of $f_{\mathbf{S}}$ are eigenvectors of the $\mathbf{B}$-pseudoinverse of the stochastic Hessian $\nabla^2 f_{\mathbf{S}}$, corresponding to eigenvalue one.
- Function f can be represented in multiple ways:

$$f(x) = \frac{1}{2}\mathrm{E}\left[\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\|^2_{\mathbf{B}}\right] = \frac{1}{2}\mathrm{E}\left[\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}}\right]. \tag{25}$$

- The gradient and Hessian of f (with respect to the $\mathbf{B}$-inner product) are given by

$$\nabla f(x) = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*), \quad \text{and} \quad \nabla^2 f = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right], \tag{26}$$

respectively, where x_* is any point in $\mathcal{L}$.

Equivalence of Algorithms

Theorem 12

Algorithm 1 (Basic Method) can be equivalently written as stochastic gradient descent (8)*, stochastic Newton method* (9)*, stochastic fixed point method* (10)*, and stochastic projection method* (11)*.*

Proof.

This follows from identities (23) in Lemma 11. □

Proof of Lemma 11 - I

Pick any $x_* \in \mathcal{L}$. First, we have

$$\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) \overset{(17)}{=} x - \mathbf{B}^{-1}\mathbf{A}^\top \mathbf{H}(\mathbf{A}x - b) \overset{(21)}{=} x - \nabla f_{\mathbf{S}}(x).$$

To establish (23), it now only remains to consider the two expressions involving the Hessian. We have

$$\nabla^2 f_{\mathbf{S}} \nabla f_{\mathbf{S}}(x) \overset{(21)+(22)}{=} \mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(19)}{=} \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \overset{(22)}{=} \nabla f_{\mathbf{S}}(x),$$

and

$$\begin{aligned}
(\nabla^2 f_{\mathbf{S}})^{\dagger_{\mathbf{B}}} \nabla f_{\mathbf{S}}(x) &\overset{(16)}{=} \mathbf{B}^{-1}(\nabla^2 f_{\mathbf{S}})^\top \left((\nabla^2 f_{\mathbf{S}})\mathbf{B}^{-1}(\nabla^2 f_{\mathbf{S}})^\top\right)^\dagger \nabla f_{\mathbf{S}}(x) \\
&\overset{(21)}{=} \mathbf{B}^{-1}(\mathbf{B}^{-1}\mathbf{Z})^\top \left((\mathbf{B}^{-1}\mathbf{Z})\mathbf{B}^{-1}(\mathbf{B}^{-1}\mathbf{Z})^\top\right)^\dagger \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&= \mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1} \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)^\dagger \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&\overset{(19)}{=} \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right) \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right)^\dagger \left(\mathbf{B}^{-1}\mathbf{Z}\mathbf{B}^{-1}\right) \mathbf{B}(x - x_*) \\
&= \mathbf{B}^{-1}\mathbf{Z}(x - x_*) \\
&\overset{(22)}{=} \nabla f_{\mathbf{S}}(x).
\end{aligned}$$

Proof of Lemma 11 - II

Identity (24) follows from

$$\frac{1}{2}\|\nabla f_{\mathbf{S}}(x)\|^2_{\mathbf{B}} \overset{(22)}{=} \frac{1}{2}(x-x_*)^\top \mathbf{Z}\mathbf{B}^{-1}\mathbf{Z}(x-x_*) \overset{(19)}{=} \frac{1}{2}(x-x_*)^\top \mathbf{Z}(x-x_*) \overset{(20)}{=} f_{\mathbf{S}}(x).$$

If $x \in \mathcal{L}$, then by picking $x_* = x$ in (22), we see that $x \in \mathcal{L}_{\mathbf{S}}$.

It remains to show that the sets defined in (i)–(iv) are identical.

- Equivalence between (i) and (ii) follows from (22).
- Now consider (ii) and (iii). Any $x_* \in \mathcal{L}$ belongs to the set defined in (iii), which follows immediately by substituting $b = \mathbf{A}x_*$. The rest follows after observing the nullspaces are identical.
- In order to show that (iii) and (iv) are equivalent, it suffices to compute $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)$ and observe that $\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x) = x$ if and only if x belongs to the set defined in (iii).

Equivalence of 4 Stochastic Reformulations

Equivalence of the Stochastic Formulations

The below theorem says that the solution sets of the fours stochastic problems (2), (4), (5), and (6) are identical. **In this sense, the four stochastic problems are equivalent.**

Theorem 13 (Equivalence of stochastic formulations)

Let $x_* \in \mathcal{L}$. The following sets are identical:

(i) $\mathcal{X} = \arg\min f(x) = \{x \ : \ f(x) = 0\} = \{x \ : \ \nabla f(x) = 0\}$ $\rightarrow$ (2)

(ii) $\mathcal{X} = \{x \ : \ \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right] \mathbf{A}x = \mathbf{B}^{-1}\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right] b\} = x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right)$ $\rightarrow$ (4)

(iii) $\mathcal{X} = \{x \ : \ \mathrm{E}\left[\Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right] = x\}$ $\rightarrow$ (5)

(iv) $\mathcal{X} = \{x \ : \ \mathrm{Prob}(x \in \mathcal{L}_{\mathbf{S}}) = 1\}$ $\rightarrow$ (6)

Moreover, $\mathcal{X}$ does not depend on $\mathbf{B}$.

Proof of Theorem 13 - Part I

As f is convex, nonnegative and achieving the value of zero (since $\mathcal{L} \neq \emptyset$), the sets in (i) are all identical. We shall now show that the sets defined in (ii)–(iv) are equal to that defined in (i).

(i) $\leftrightarrow$ (ii): Using the formula for the gradient from (26), we see that

$$\begin{aligned} \{x \ : \ \nabla f(x) = 0\} &= \{x \ : \ \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right](x - x_*) = 0\} \\ &= \{x \ : \ \mathrm{E}\left[\mathbf{Z}\right](x - x_*) = 0\} \\ &= x_* + \{h \ : \ \mathrm{E}\left[\mathbf{Z}\right] h = 0\} \\ &= x_* + \mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right), \end{aligned}$$

which shows that (i) and (ii) are the same.

(i) $\leftrightarrow$ (iii): Equivalence of (i) and (iii) follows by taking expectations in (23) to obtain

$$\nabla f(x) = \mathrm{E}\left[\nabla f_{\mathbf{S}}(x)\right] \overset{(23)}{=} \mathrm{E}\left[x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right].$$

Proof of Theorem 13 - Part II

$(i) \leftrightarrow (iv)$: It remains to establish equivalence between (i) and (iv). Let

$$\mathcal{X} = \{x \ : \ f(x) = 0\} \overset{(25)}{=} \left\{x \ : \ \mathrm{E}\left[\left\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right\|^2_{\mathbf{B}}\right] = 0\right\} \tag{27}$$

and let $\mathcal{X}'$ be the set from (iv).

We need to show that $\mathcal{X}' = \mathcal{X}$. For easier reference, let

$$\xi_{\mathbf{S}}(x) \overset{\text{def}}{=} \left\|x - \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right\|^2_{\mathbf{B}}.$$

Note that the following three probabilistic events are identical:

$$[x \in \mathcal{L}_{\mathbf{S}}] = \left[x = \Pi^{\mathbf{B}}_{\mathcal{L}_{\mathbf{S}}}(x)\right] = [\xi_{\mathbf{S}}(x) = 0]. \tag{28}$$

We first show that $\mathcal{X}' \subseteq \mathcal{X}$.

In view of (28), if $x \in \mathcal{X}'$, then the random variable $\xi_{\mathbf{S}}(x)$ is equal to zero with probability 1, which implies $\mathrm{E}[\xi_{\mathbf{S}}(x)] = 0$, whence $x \in \mathcal{X}$.

Proof of Theorem 13 - Part III

Let us now sow that $\mathcal{X} \subseteq \mathcal{X}'$.

Let $1_{[\xi_{\mathbf{S}}(x) \geq t]}$ be the indicator function of the event $[\xi_{\mathbf{S}}(x) \geq t]$. Note that since $\xi_{\mathbf{S}}(x)$ is a nonnegative random variable, for all $t \in \mathbb{R}$ we have the inequality

$$\xi_{\mathbf{S}}(x) \geq t 1_{\xi_{\mathbf{S}}(x) \geq t}. \tag{29}$$

Now take $x \in \mathcal{X}$ and consider $t > 0$. By taking expectations in (29), we obtain

$$0 = \mathrm{E}[\xi_{\mathbf{S}}(x)] \geq \mathrm{E}\left[t 1_{\xi_{\mathbf{S}}(x) \geq t}\right] = t\mathrm{E}\left[1_{\xi_{\mathbf{S}}(x) \geq t}\right] = t\mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t),$$

which implies that $\mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t) = 0$. Now choose $t_i = 1/i$ for $i = 1, 2, \ldots$ and note that the event $[\xi_{\mathbf{S}}(x) > 0]$ can be written as

$$[\xi_{\mathbf{S}}(x) > 0] = \bigcup_{i=1}^{\infty} [\xi_{\mathbf{S}}(x) \geq t_i].$$

Proof of Theorem 13 - Part IV

Therefore, by the union bound,

$$\mathrm{Prob}(\xi_{\mathbf{S}}(x) > 0) \leq \sum_{i=1}^{\infty} \mathrm{Prob}(\xi_{\mathbf{S}}(x) \geq t_i) = 0,$$

which immediately implies that $\mathrm{Prob}(\xi_{\mathbf{S}}(x) = 0) = 1$. From (28) we conclude that $x \in \mathcal{X}'$.

Independence on B. Since characterization (iv) of $\mathcal{X}$ does not depend on $\mathbf{B}$, we conclude that $\mathcal{X}$ does not depend on $\mathbf{B}$.

Exactness of the Reformulations

Rangespace and Nullspace of a Matrix - I

Let $\mathbf{M} \in \mathbb{R}^{m \times n}$.

Definition 14 (Rangespace of a matrix)

By $\mathrm{Range}\,(\mathbf{M})$ we mean the **rangespace of matrix M.** This is the linear subspace of $\mathbb{R}^m$ generated by the columns of $\mathbf{M}$:

$$\mathrm{Range}\,(\mathbf{M}) \stackrel{\text{def}}{=} \{\mathbf{M}x \ : \ x \in \mathbb{R}^n\} = \left\{\sum_j \mathbf{M}_{:j} x_j, \quad x \in \mathbb{R}^n\right\}.$$

Definition 15 (Nullspace of a matrix)

By $\mathrm{Null}\,(\mathbf{M})$ we mean the **nullspace of matrix M.** This is the linear subspace of $\mathbb{R}^n$ formed by the vectors orthogonal (under standard Euclidean inner product) to all rows of $\mathbf{M}$:

$$\mathrm{Null}\,(\mathbf{M}) \stackrel{\text{def}}{=} \{x \in \mathbb{R}^n \ : \ \mathbf{M}x = 0\} = \left\{x \in \mathbb{R}^n \ : \ \langle \mathbf{M}_{i:}^\top, x\rangle = 0 \quad \forall i\right\}.$$

Rangespace and Nullspace of a Matrix - II

Definition 16 (Orthogonal complement)

Let X be a subspace of a vector space. The **orthogonal complement** of X is the linear subspace $X^\perp \stackrel{\text{def}}{=} \{y \ : \ \langle y, x\rangle = 0 \ \forall x \in X\}$.

Here we collect some useful identities involving rangespaces and nullspaces of a matrix:

Fact 17

For any $\mathbf{M} \in \mathbb{R}^{m \times n}$, *we have*

(i) $\mathrm{Range}\left(\mathbf{M}^\top\right) = \mathrm{Null}\,(\mathbf{M})^\perp$

(ii) $\mathrm{Range}\left(\mathbf{M}^\top\right)^\perp = \mathrm{Null}\,(\mathbf{M})$

(iii) *If* $\mathbf{G} \succ 0$, *then* $\mathrm{Null}\left(\mathbf{M}^\top \mathbf{G} \mathbf{M}\right) = \mathrm{Null}\,(\mathbf{M})$

Exactness

Key Question: When are the stochastic formulations (2), (4), (5), (6) equivalent to the linear system (1)? That is, when is their set of solutions $\mathcal{X}$ identical to solution set of the linear system $\mathcal{L}$?

This leads to the concept of **exactness:**

Assumption 3 (Exactness)

Stochastic reformulations (2), (4), (5), (6) *of problem* (1) *are* exact. *That is,* $\mathcal{X} = \mathcal{L}$.

In what follows, we will

- Give **sufficient,** and **necessary & sufficient** conditions for exactness.
- Use this assumption to **prove convergence** of the algorithms to a specific point in $\mathcal{L}$.

Necessary and Sufficient Conditions for Exactness

Theorem 18 ($\Leftrightarrow$ Conditions for exactness)

The following statements are equivalent:

(i) *Assumption 3 (exactness) holds*

(ii) $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{A}\right)$

(iii) $\mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}\right) = \mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right)$

(iv) $\mathrm{Range}\left(\mathbf{A}\right) \cap \mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right) = \{0\}$

Proof of Theorem 18 - I

(i) $\leftrightarrow$ *(ii)*: Choose any $x_* \in \mathcal{L}$. We know that $\mathcal{L} = x_* + \mathrm{Null}\,(\mathbf{A})$. On the other hand, Theorem 13 says that $\mathcal{X} = x_* + \mathrm{Null}\,(\mathrm{E}\,[\mathbf{Z}])$.

(ii) $\leftrightarrow$ *(iii)*: If (ii) holds, then

$$\mathrm{Null}\,(\mathbf{A}) = \mathrm{Null}\,(\mathrm{E}\,[\mathbf{Z}]) = \mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\,[\mathbf{Z}]\right),$$

and (iii) follows. If (iii) holds, then

$$\mathrm{Null}\,(\mathbf{A}) = \mathrm{Null}\left(\mathbf{B}^{-1/2}\mathrm{E}\,[\mathbf{Z}]\right) = \mathrm{Null}\,(\mathrm{E}\,[\mathbf{Z}]),$$

proving (ii).

Proof of Theorem 18 - II

(ii) $\leftrightarrow$ *(iv)*: First, note that

$$\mathrm{E}\,[\mathbf{Z}] = \mathbf{A}^\top(\mathrm{E}\,[\mathbf{H}])^{1/2}(\mathrm{E}\,[\mathbf{H}])^{1/2}\mathbf{A}.$$

In view of Fact 17, for any matrix $\mathbf{M}$ we have $\mathrm{Null}\left(\mathbf{M}^\top\mathbf{M}\right) = \mathrm{Null}\,(\mathbf{M})$. Therefore,

$$\mathrm{Null}\,(\mathrm{E}\,[\mathbf{Z}]) = \mathrm{Null}\left((\mathrm{E}\,[\mathbf{H}])^{1/2}\mathbf{A}\right).$$

Moreover, we know that

(a) $\mathrm{Null}\left((\mathrm{E}\,[\mathbf{H}])^{1/2}\mathbf{A}\right) = \mathrm{Null}\,(\mathbf{A})$ if and only if $\mathrm{Range}\,(\mathbf{A}) \cap \mathrm{Null}\left((\mathrm{E}\,[\mathbf{H}])^{1/2}\right) = \{0\}$, and

(b) $\mathrm{Null}\left((\mathrm{E}\,[\mathbf{H}])^{1/2}\right) = \mathrm{Null}\,(\mathrm{E}\,[\mathbf{H}])$ (see Fact 17).

It remains to combine these observations.

Sufficient Conditions for Exactness

We now list some sufficient conditions for exactness.

Lemma 19 (Sufficient conditions for exactness)

Any of these conditions implies that Assumption 3 is satisfied:

(i) $\mathrm{E}\left[\mathbf{H}\right] \succ 0$

(ii) $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{H}\right]\right) \subseteq \mathrm{Null}\left(\mathbf{A}^\top\right)$

Proof.

If (i) holds, then $\mathrm{Null}\left(\mathrm{E}\left[\mathbf{Z}\right]\right) = \mathrm{Null}\left(\mathbf{A}^\top \mathrm{E}\left[\mathbf{H}\right] \mathbf{A}\right) = \mathrm{Null}\left(\mathbf{A}\right)$, where the last equality follows from Fact 17. Exactness now follows by applying Theorem 18.

On the other hand, in view of Fact 17, (ii) implies statement (iv) in Theorem 18, and hence exactness follows. □

Condition Number

Spectral Decomposition of the Hessian of f

Recall that the **Hessian of f** is given by

$$\nabla^2 f = \mathrm{E}_{\mathbf{S}\sim\mathcal{D}}\left[\nabla^2 f_{\mathbf{S}}\right] = \mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]. \tag{30}$$

Lemma 20
Matrices $\mathbf{B}^{-1}\mathrm{E}\left[\mathbf{Z}\right]$ and $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$ have the same eigenvalues.

Proof.
It is known that for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$, the matrices $\mathbf{XY}$ and $\mathbf{YX}$ have the same eigenvalues. It only remains to apply this to $\mathbf{X} = \mathbf{B}^{-1/2}$ and $\mathbf{Y} = \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]$. □

The above result allows us to study spectral properties of the Hessian $\nabla^2 f$ through the **eigenvalue decomposition** of the symmetric positive definite matrix $\mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2}$.

Eigenvalues of the Hessian of f

Let

$$\mathbf{W} \overset{\text{def}}{=} \mathbf{B}^{-1/2}\mathrm{E}\left[\mathbf{Z}\right]\mathbf{B}^{-1/2} = \mathbf{U}\Lambda\mathbf{U}^\top = \sum_{i=1}^{n} \lambda_i u_i u_i^\top \tag{31}$$

be the **eigenvalue decomposition of W,** where

$$\mathbf{U} = [u_1, \ldots, u_n] \in \mathbb{R}^{n\times n}$$

is an orthonormal matrix composed of **eigenvectors** (i.e., we have $\mathbf{U}\mathbf{U}^\top = \mathbf{U}^\top\mathbf{U} = \mathbf{I}$), and

$$\Lambda = \mathbf{Diag}\left(\lambda_1, \lambda_2, \ldots, \lambda_n\right)$$

is a diagonal matrix of **eigenvalues.** Assume without loss of generality that the eigenvalues are ordered from largest to smallest:

$$\lambda_{\max} \overset{\text{def}}{=} \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \overset{\text{def}}{=} \lambda_{\min}.$$

All Eigenvalues of W are Between 0 and 1

Lemma 21
$0 \leq \lambda_i \leq 1$ *for all* i.

Proof.
Since $\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}$ is symmetric positive semidefinite, so is its expectation $\mathbf{W}$, implying that $\lambda_i \geq 0$ for all i.

Further, note that $\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}$ is a projection matrix. Indeed, it is the projection (in the standard $\mathbf{I}$-norm) onto $\mathrm{Range}\left(\mathbf{B}^{-1/2}\mathbf{A}^\top\mathbf{S}\right)$. Therefore, its eigenvalues are all zeros or ones. Since the map $\mathbf{X} \mapsto \lambda_{\max}(\mathbf{X})$ is convex, by **Jensen's inequality** we get

$$\lambda_{\max}(\mathbf{W}) = \lambda_{\max}\left(\mathrm{E}\left[\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2}\right]\right) \leq \mathrm{E}\left[\lambda_{\max}(\mathbf{B}^{-1/2}\mathbf{Z}\mathbf{B}^{-1/2})\right] \leq 1.$$

□

Smallest Nonzero Eigenvalue

Lemma 22
If Assumption 3 (exactness) holds, then $\lambda_{\max} > 0$.

Proof.
Assume, by contradiction, that $\lambda_i = 0$ for all i. Then from Theorem 18 and the fact that $\mathrm{Null}\left(\mathbf{W}\right) = \mathrm{Range}\left(u_i \ : \ \lambda_i = 0\right)$ we conclude that $\mathrm{Null}\left(\mathbf{A}\mathbf{B}^{-1/2}\right) = \mathbb{R}^n$, which in turn implies that $\mathrm{Null}\left(\mathbf{A}\right) = \mathbb{R}^n$. This can only happen if $\mathbf{A} = 0$, which is contradicts with our assumption on $\mathbf{A}$. □

Now, let j be the largest index for which $\lambda_j > 0$. This identifies the **smallest nonzero eigenvalue of W,** which we shall denote as

$$\lambda^+_{\min} = \lambda_j.$$

If all eigenvalues $\{\lambda_i\}$ are positive, then $j = n$.

Condition Number

Definition 23

The **condition number** associated with the four stochastic reformulations is the quantity[3]

$$\zeta(\mathbf{A}, \mathbf{B}, \mathcal{D}) = \zeta \stackrel{\text{def}}{=} \|\mathbf{W}\|\|\mathbf{W}^\dagger\| = \frac{\lambda_{\max}}{\lambda_{\min}^+}. \tag{32}$$

Remark:

- As we shall see, convergence rate of the Basic Method is described by ζ.
- As one varies the parameters defining the reformulation (i.e., $\mathcal{D}$ and $\mathbf{B}$), ζ changes. As a general rule of thumb, simple distributions will lead to reformulations with a small condition number. For instance, choosing $\mathbf{S} = \mathbf{I}$ with probability one gives $\zeta = 1$. However, in such a case each step of the Basic Method is very expensive. One needs to strike the right balance.

[3] $\|\mathbf{X}\|$ denotes the **spectral norm** of $\mathbf{X}$. In general, $\|\mathbf{X}\| = \left(\lambda_{\max}(\mathbf{X}^\top\mathbf{X})\right)^{1/2}$. If $\mathbf{X}$ is symmetric positive semidefinite, then $\|\mathbf{X}\|^2 = \lambda_{\max}(\mathbf{X}^\top\mathbf{X}) = \lambda_{\max}(\mathbf{X}^2) = (\lambda_{\max}(\mathbf{X}))^2$. Therefore, $\|\mathbf{X}\| = \lambda_{\max}(\mathbf{X})$.