SIAM J. Optim.
Vol. 25, No. 1, pp. 234–260


# PROBABILISTIC INTERPRETATION OF LINEAR SOLVERS*

PHILIPP HENNIG†

**Abstract.** This paper proposes a probabilistic framework for algorithms that iteratively solve unconstrained linear problems $Bx = b$ with positive definite $B$ for $x$. The goal is to replace the point estimates returned by existing methods with a Gaussian posterior belief over the elements of the inverse of $B$, which can be used to estimate errors. Recent probabilistic interpretations of the secant family of quasi-Newton optimization algorithms are extended. Combined with properties of the conjugate gradient algorithm, this leads to uncertainty-calibrated methods with very limited cost overhead over conjugate gradients, a self-contained novel interpretation of the quasi-Newton and conjugate gradient algorithms, and a foundation for new nonlinear optimization methods.

**Key words.** linear programming, quasi-Newton methods, conjugate gradient, Gaussian inference

**AMS subject classifications.** 49M15, 65K05, 60G15

**DOI.** 10.1137/140955501

## 1. Introduction.

### 1.1. Motivation.

Solving the unconstrained linear problem of finding $x$ in

$$Bx = b \qquad \text{with symmetric, positive definite } B \in \mathbb{R}^{N\times N} \text{ and } b \in \mathbb{R}^N \tag{1.1}$$

is a basic task for computational linear algebra. It is equivalent to minimizing the quadratic $f(x) = {}^1\!/_2 x^\intercal Bx - x^\intercal b$, with gradient $F(x) = \nabla_x f(x) = Bx - b$ and constant Hessian $B$. If $N$ is too large for exact solution, iterative solvers such as the method of conjugate gradients [27] (CG) are widely applied. These methods produce a sequence of estimates $\{x_i\}_{i=0,\dots,M}$, updated by evaluating $F(x_i)$. The question addressed here is: Assume we run an iterative solver for $M < N$ steps. How much information does doing so provide about $B$ and its (pseudo-)inverse $H$? If we had to give estimates for $B$, $H$, and for the solution to related problems $B\tilde{x} = \tilde{b}$, what should they be, and how big of an "error bar" (a joint posterior distribution) should we put on these estimates? The gradient $F(x_i)$ provides an error residual on $x_i$, but not on $B, H$, and $\tilde{x}$.

It will turn out that a family of quasi-Newton methods (section 1.2), more widely used to solve nonlinear optimization problems, can help answer this question, because classic derivations of these methods can be reformulated and extended into a probabilistic interpretation of these methods as maxima of Gaussian posterior probability distributions (section 2). The covariance of these Gaussians offers a new object of interest and provides error estimates (section 3). Because there are entire linear spaces of Gaussian distributions with the same posterior mean but differing posterior error estimates, selecting one error measure consistent with the algorithm is a new statistical estimation task (section 4).

### 1.2. The Dennis family of secant methods.

The family of *secant* update rules for an approximation to the Newton–Raphson search direction is among the most

*Received by the editors February 4, 2014; accepted for publication (in revised form) October 13, 2014; published electronically January 20, 2015.
http://www.siam.org/journals/siopt/25-1/95550.html
†Max Planck Institute for Intelligent Systems, 72076 Tübingen, Germany (phennig@tuebingen.mpg.de).

popular building blocks for continuous nonlinear programming. Their evolution chiefly occurred from the late 1950s [8] to the 1970s, and is widely understood to be crowned by the development of the BFGS rule due to Broyden [5], Fletcher [18], Goldfarb [22], and Shanno [39], which now forms a core part of many contemporary optimization methods. But the family also includes the earlier and somewhat less popular DFP rule of Davidon [8] and Fletcher and Powell [19], the Greenstadt [23] rule, and the so-called symmetric rank-1 method [8, 4]. Several authors have proposed grouping these methods into broader classes, among them Broyden in 1967 [4] (subsequently refined by Fletcher [18]) and Davidon in 1975 [9]. Of particular interest here will be a class of updates formulated in 1971 by Dennis [10], which includes all of the specific rules cited above. It is the class of update rules mapping a current estimate $B_0$ for the Hessian, a vector-valued pair of observations $y_i = F(x_i) - F(x_{i-1}) \in \mathbb{R}^N$ and $s_i = x_i - x_{i-1}$ with $y_i = Bs_i$, into a new estimate $B_i$ of the form

(1.2)
$$B_{i+1} = B_i + \frac{(y_i - B_i s_i)c_i^\intercal + c_i(y_i - B_i s_i)^\intercal}{c_i^\intercal s_i} - \frac{c_i s_i^\intercal (y_i - B_i s_i) c_i^\intercal}{(c_i^\intercal s_i)^2} \qquad \text{for some } c_i \in \mathbb{R}^N.$$

This ensures the secant relation $y_i = B_{i+1}s_i$, sometimes called "the quasi-Newton equation" [13]. Convergence of the sequence of iterates $x_i$ for various members of this class (and the classes of Broyden and Davidon) are well-understood [21, 12]. The rules named above can be found in the Dennis class as [30]

| | | |
|---|---|---|
| (1.3) | symmetric rank-1 (SR1) | $c = y - B_0 s,$ |
| (1.4) | Powell symmetric Broyden [37] | $c = s,$ |
| (1.5) | Greenstadt [23] | $c = B_0 s,$ |
| (1.6) | Davidon Fletcher Powell (DFP) | $c = y,$ |
| (1.7) | Broyden Fletcher Goldfarb Shanno (BFGS) | $c = y + \sqrt{\frac{y^\intercal s}{s^\intercal B_0 s}} B_0 s.$ |

*Inverse updates.* Because the update of (1.2) is of rank 2, the corresponding estimate for the inverse $H = B^{-1}$ (assuming it exists) can be constructed using the matrix inversion lemma. Alternatively, all Dennis rules can also be used as *inverse updates* [13], i.e., estimates for $H$ itself, by exchanging $s \leftrightarrow y$ and $B \leftrightarrow H$, $B_0 \leftrightarrow H_0$ above (corresponding to the secant relation $s = Hy$). Interestingly, the DFP and BFGS updates are duals of each other under this exchange [13]: The inverse of $B_1$ as constructed by the DFP rule (1.6) equals the $H_1$ arising from the inverse BFGS rule (1.7). This does not mean BFGS and DFP are the same, but only that they fill opposing roles in the inverse and direct formulation. To avoid confusion, in this text the DFP rule will always be used in the sense of a direct update (estimating $B$, with $c = y$), and the BFGS rule in the inverse sense (i.e., estimating $H$, with $c = s$). The first parts of this text will focus on direct updates and thus mostly talk about the DFP method instead of the BFGS rule. All results extend to the inverse models (and thus BFGS) under the exchange of variables mentioned above. Sections 3.2 and 4 will make some specific choices geared to inverse updates. They will then talk explicitly about BFGS, always in the sense of an inverse update.

*Towards probabilistic quasi-Newton methods.* This text gives a probabilistic interpretation of the Dennis family for the linear problems of (1.1). We will interpret the secant methods as *estimators* of (inverse) Hessians of an objective function, and ask what kind of prior assumptions would give rise to these specific estimators. This re-

sults in a self-contained derivation of inference rules for symmetric matrices. Some of the rules quoted above can be motivated as "natural" from the inference perspective.

Another major strand of nonlinear optimization methods extends from the conjugate gradient algorithm of Hestenes and Stiefel [27] for linear problems, nonlinearly extended by Fletcher and Reeves [20] and others. On linear problems, the CG and quasi-Newton ideas are closely linked: Nazareth [33] showed that CG is equivalent to BFGS for linear problems (with exact line searches, when the initial estimate $B_0 = \boldsymbol{I}$). More generally, Dixon [16, 17] showed for quasi-Newton methods in the Broyden class (which also contains the methods listed above) that, under exact line searches and the same starting point, all methods in Broyden's class generate a sequence of points identical to CG, if the starting matrix $B_0$ is taken as a preconditioner of CG. In this sense, this text also provides a novel derivation for conjugate gradients, and will use several well-known properties of that method. Implications of the results presented herein to nonlinear variants of conjugate gradients will be left for future work.

**1.3. Numerical methods perform inference—the value of a statistical interpretation.** The defining aspect of quasi-Newton methods is that they approximate—*estimate*—the Hessian matrix of the objective function, or its inverse, based on evaluations—*observations*—of the objective's gradient and certain *prior* structural restrictions on the estimate. They can therefore be interpreted as inferring the latent quantity $B$ or $H$ from the observed quantities $s, y$. This creates a connection to statistics and probability theory, in particular the probabilistic framework of encoding prior assumptions in a probability measure over a hypothesis space, and describing observations using a likelihood function, which combines with the prior according to Bayes' theorem into a posterior measure over the hypothesis space (section 2).

On the one hand, this elucidates prior assumptions of quasi-Newton methods (section 3). On the other hand, it suggests new functionality for the existing methods, in particular error estimates on $B$ and $H$ (section 4). In future work, it may also allow for algorithms robust to "noisy" linear maps, such as they arise in physical inverse problems.

The interpretation of numerical problems as estimation was pointed out by statisticians like Diaconis in 1988 [14] and O'Hagan in 1992 [36], well after the introduction of quasi-Newton methods. To the best of the author's knowledge, the idea has rarely attracted interest in numerical mathematics, and has not been studied in the context of quasi-Newton methods before the recent work by Hennig and Kiefel [25, 26]. An argument sometimes raised against analyzing numerical methods probabilistically is that numerical problems do not generally feature an aspect of randomness. But probability theory makes no formal distinction between epistemic uncertainty, arising from lack of knowledge, and aleatoric uncertainty, arising from "randomness," whatever the latter may be taken to mean precisely. Randomness is not a prerequisite for the use of probabilities. Those who do feel uneasy about applying probability theory to unknown deterministic quantities, however, may prefer another, perhaps more subjective argument: From the point of view of a numerical algorithm's designer, the "population" of problems that practitioners will apply the algorithm to does, in fact, form a probability distribution from which tasks are "sampled."

Numerical algorithms running on a finite computational budget make numerical errors. A notion of the imprecision of these answers is helpful, in particular when the method is used within a larger computational framework. Explicit error estimates can be propagated through the computational pipeline, helping identify points of instability, and to distribute or save computational resources. Needless to say, it

makes no sense to ask for the *exact* error (if the exact difference between the true and estimated answer were known, the exact answer would be known, too). But it is meaningful to ask for the remaining volume of hypotheses consistent with the computations so far. This paper attempts to construct such an answer for linear problems.

**1.4. Overview of main results.** As pointed out above, although quasi-Newton methods are most popular for nonlinear optimization, here the focus will be on *linear* problems. Extending the probabilistic interpretation constructed here to the nonlinear setting of inferring the (inverse) Hessian of a function $f$ will be left for future work (see [26] for pointers). The present aim is an iterative linear solver iterating through posterior beliefs $\{p_t(x, H)\}_{t=1,\dots,M}$ for $H = B^{-1}$ and the solution $x = Hb$ of the linear problem. These beliefs will be constructed as Gaussian densities $p_t(H) = \mathcal{N}(H; H_t, V_t)$ over the elements[1] of $H$, with mean $H_t$ and covariance matrix $V_t$.

The results in this paper significantly clarify and extend previous results by Hennig and Kiefel [26] and Hennig [24]. Here is a brief look at the main results.

**Dennis family derived in a symmetric hypothesis class (section 2).** As a probabilistic interpretation of results by Dennis and Moré [13] and Dennis and Schnabel [11], Hennig [24] provided a derivation of rank-2 secant methods in terms of two independent observations of two separate parts of the Hessian. This viewpoint affords a nonparametric extension to nonlinear optimization, but is not particularly elegant. This paper provides a cleaner derivation: the Dennis family can, in fact, be derived naturally from a prior over only symmetric matrices. This extends the results of Dennis and Schnabel [11], from statements about the maximum of a Frobenius norm in the space of symmetric matrices to the entire structure of that norm in that space.

**Interpretation of SR1, Greenstadt, DFP, and BFGS (section 3).** The choice of prior parameters distinguishes between the members of the Dennis family. An analysis shows that DFP and BFGS are "more correct" than other members of the family in the sense that they are consistent with exact probabilistic inference for the entire run of the algorithm, while general Dennis rules are only consistent after the first step (Lemmas 3.2 and 3.3). Further, SR1, Greenstadt, DFP, and BFGS all use different prior measures that, although all "scale-free," give imperfect notions of calibration for the prior measure. Finally, because BFGS is equivalent to CG (see [33] and Corollary 3.4 below), its set of evaluated gradients is orthogonal. This allows a computationally convenient parameterization of posterior uncertainty. Overall, the picture arising is that, from the probabilistic perspective, the DFP and particularly BFGS methods have convenient numerical properties, but their posterior measure can be calibrated better.

**Posterior uncertainty by parameter estimation (section 4).** It will transpire that the decision for a specific member of the Dennis family still leaves a space of possible choices of prior covariances consistent with this update rule. Constructing a meaningful posterior uncertainty estimate (covariance) on $H$ after finitely many steps requires a choice in this unidentified space, which, as in other estimation problems, needs to be motivated based on some notion of regularity in $H$. Several possible choices are discussed in section 3, all of

[1] For notational convenience, the elements of $H$ will be treated as the elements of a vector of length $N^2$; see the beginning of section 2.2.

which add very low overhead to the standard conjugate gradient algorithm.

## 2. Gaussian inference from matrix-vector multiplications.

**2.1. Introduction to Gaussian inference.** Gaussian inference—probabilistic inference using both a Gaussian prior and a Gaussian likelihood—is one of the best-studied areas of probabilistic inference. The following is a very brief introduction; more can be found in introductory texts [38, 29]. Consider a hypothesis class consisting of elements of the $D$-dimensional real vector space, $v \in \mathbb{R}^D$, and assign a Gaussian prior probability density over this space:

$$p(v) = \mathcal{N}(v; \mu, \Sigma) := \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(v-\mu)^\intercal \Sigma^{-1}(v-\mu)\right), \tag{2.1}$$

parametrized by *mean vector* $\mu \in \mathbb{R}^D$ and positive definite *covariance matrix* $\Sigma \in \mathbb{R}^{D\times D}$. If we now observe a linear mapping $A^\intercal v + a = y \in \mathbb{R}^M$ of $v$, up to Gaussian uncertainty of covariance $\Lambda \in \mathbb{R}^{M\times M}$, i.e., according to the likelihood function

$$p(y \mid A, a, v) = \mathcal{N}(y; A^\intercal v + a, \Lambda), \tag{2.2}$$

then, by Bayes' theorem and a simple linear computation (see, e.g., [38, section 2.1.2]), the posterior, the unique distribution over $v$ consistent with both prior and likelihood, is

$$p(v \mid y, A, a) = \mathcal{N}[v; \mu + \Sigma A(A^\intercal \Sigma A + \Lambda)^{-1}(y - A^\intercal \mu - a), \Sigma - \Sigma A(A^\intercal \Sigma A + \Lambda)^{-1} A^\intercal \Sigma]. \tag{2.3}$$

This derivation also works in the limit of perfect information, i.e., for a well-defined limit of $\Lambda \to 0$, in which case[2] the likelihood converges to the Dirac distribution $p(y \mid A, a, v) \to \delta(y - A^\intercal v - a)$. The crucial point is that constructing the posterior after linear observations involves only linear algebraic operations, with the posterior covariance (the "error bar") using many of the computations also required to compute the mean (the "best guess").

Gaussian inference is closely linked to least-squares estimation: Because the logarithm is concave, the maximum of the posterior (2.3) (which equals the mean) is also the minimizer of the quadratic norm (using $\|x\|_K^2 := x^\intercal K^{-1} x$)

$$-2\log p(v \mid y, A, a) = \|y - A^\intercal v - a\|_\Lambda^2 + \|v - \mu\|_\Sigma^2 + \text{const}\,. \tag{2.4}$$

The added value of the probabilistic interpretation is embodied in the posterior covariance, which quantifies remaining degrees of freedom of the estimator and can thus also be interpreted as a measure of uncertainty, or estimated error.

**2.2. Inference on asymmetric matrices from matrix vector multiplications.** We now consider Gaussian inference in the specific context of iterative solvers for linear problems as defined in (1.1). Our solver shall maintain a current probability density estimate, either $p_i(B)$ or $p_i(H)$, $i = 0, \ldots, M$. The solver does not have direct access to $B$ itself, but only to a function mapping $s \to Bs$, for arbitrary $s \in \mathbb{R}^N$.

It is possible to use the Gaussian inference framework in the context of secant methods [26] through the use of Kronecker algebra: We write the elements of $B$ as

[2] If $A$ is not of maximal rank, a precise formulation requires a projection of $y$ into the preimage of $A$. This is merely a technical complication. It is circumvented here by assuming, later on, that line-search directions are linearly independent. This amounts to a maximal-rank $A$.

a vector $\overrightarrow{B} \in \mathbb{R}^{N^2}$, indexed as $\overrightarrow{B}_{ij}$ by the matrix' index set[3] $(i,j) \in \mathbb{R} \times \mathbb{R}$. The Kronecker product provides the link between such "vectorized matrices" and linear operations (e.g., [40]). The Kronecker product $A \otimes C$ of two matrices $A \in \mathbb{R}^{M_a \times N}$ and $C \in \mathbb{R}^{M_c \times N}$ is the $M_a M_c \times N^2$ matrix with elements $(A \otimes C)_{(ij)(k\ell)} = A_{ik} C_{j\ell}$. It has the property $(A \otimes C)\overrightarrow{B} = \overrightarrow{ABC^\intercal}$. Thus, $\overrightarrow{BS}$ can be written as $(\boldsymbol{I} \otimes S)\overrightarrow{B}$, which allows incorporating the kind of observations made by an iterative solver in a Gaussian inference framework, according to the following lemma.

LEMMA 2.1 (proof in Hennig and Kiefel, 2013 [26]). *Given a Gaussian prior over a general quadratic matrix $\overrightarrow{B}$, with prior mean $\overrightarrow{B}_0$ and a prior covariance with Kronecker structure, $p(B) = \mathcal{N}(\overrightarrow{B}; \overrightarrow{B}_0, W \otimes W)$, the posterior mean after observing $BS = Y \in \mathbb{R}^{N \times M}$ (i.e., $M$ projections along the line-search directions $S \in \mathbb{R}^{N \times M}$) is*

$$B_M = B_0 + (Y - B_0 S)(S^\intercal W S)^{-1} W S^\intercal \tag{2.5}$$

*and the posterior covariance is*

$$V_M = W \otimes \left[W - WS(S^\intercal W S)^{-1} W S^\intercal\right]. \tag{2.6}$$

This implies, for example, that Broyden's rank-1 method [3] is equal to the posterior mean update after a single line search for the parameter choice $W = \boldsymbol{I}$. This is a probabilistic rephrasing of the much older observation, most likely by Dennis and Moré [13], that Broyden's method minimizes a change in the Frobenius norm $\|B_i - B_{i-1}\|_{F,\boldsymbol{I}}$ such that $B_i s_i = y_i$. The weighted Frobenius norm $\|A\|^2_{F,W} = \operatorname{tr}(AW^{-1}A^\intercal W^{-1})$ (with the positive definite weighting $W$) is the $\ell_2$ loss on vectorized matrices in the sense that $\|A\|^2_{F,W} = \|\overrightarrow{A}\|^2_{W \otimes W}$.

An important observation is that Broyden's method ceases to be a direct match to this update after the first line search, because matrix $S^\intercal W S$ is not a diagonal matrix. This matrix will come to play a central role; we will call it *the Gram matrix*, because it is an inner product of $S$ weighted by the positive definite $W$.

**2.2.1. Symmetric hypothesis classes.** It is well known that, because the posterior mean of (2.5) is not in general a symmetric matrix, it is a suboptimal learning rule for the Hessian of an objective function, which is why this class was quickly abandoned in favor of the rank-2 updates in the Dennis family mentioned above. Dennis and Moré [13] and Dennis and Schnabel [11] showed that the minimizer of weighted Frobenius regularizers (the maximizer of the Gaussian posterior) within the linear subspace of symmetric matrices is given by the Dennis class of update rules. Hennig and Kiefel [26] constructed a probabilistic interpretation based on this derivation, which involves doubling the input domain of the objective function and introducing two separate, independent observations. This has the advantage of allowing for relatively straightforward nonparametric extensions, and a broad class of noise models for cases in which gradients cannot be evaluated without error [24]. But artificially doubling the input dimensionality is dissatisfying.

We now introduce a cleaner derivation of the same updates by explicitly restricting the hypothesis class to symmetric matrices. This gives the covariance matrix a more involved structure than the Kronecker product, and makes derivations more challenging. It results in a new interpretation for the Dennis class, fully consistent

[3] In the notation used here, this vector is assumed to be created by stacking the elements of $B$ row after row into a column vector. An equivalent column-by-column formulation is also widely used. In that formulation, some of the formulae below are permuted.

with the probabilistic framework. Since the identity of the posterior mean was known from [13, 11, 26], the interesting novel aspect here is the structure of the posterior covariance. In essence, it provides insight into the structure of the loss function *around* the previously known estimates.

We begin by building a Gaussian prior over the symmetric matrices, using the symmetrization operator $\Gamma$, the linear operator acting on vectorized matrices defined implicitly through its effect $\Gamma\overrightarrow{A} = \frac{1}{2}\overrightarrow{(A + A^\intercal)}$ (see the explicit definition in Appendix A.1).

LEMMA 2.2 (proof in Appendix A.1). *Assuming a Gaussian prior* $p(B) = \mathcal{N}(\overrightarrow{B}; \overrightarrow{B}_0, W \otimes W)$ *over the space of square matrices* $B \in \mathbb{R}^{N\times N}$ *with Kronecker covariance* $\mathrm{cov}(B_{ij}, B_{k\ell}) = W_{ik}W_{j\ell}$ *(this requires* $W$ *to be a symmetric positive definite matrix), the prior over the symmetric matrix* $\Gamma\overrightarrow{B}$ *is* $p(B) = \mathcal{N}(\Gamma\overrightarrow{B}; \Gamma\overrightarrow{B}_0, W \mathbin{\overline{\otimes}} W)$.

Here, $W \mathbin{\overline{\otimes}} W = \Gamma(W \otimes W)\Gamma^\intercal$ is the *symmetric Kronecker product* of $W$ with itself (see, e.g., [40] for an earlier mention). It is the matrix containing elements

$$(W \mathbin{\overline{\otimes}} W)_{ij,k\ell} = \tfrac{1}{2}(W_{ik}W_{j\ell} + W_{jk}W_{i\ell}). \tag{2.7}$$

It can easily be seen that, when acting on a square (not necessarily symmetric) vectorized matrix $K \in \mathbb{R}^{N\times N}$, it has the effect $(W \mathbin{\overline{\otimes}} W)\overrightarrow{K} = \frac{1}{2}(WKW^\intercal + W^\intercal K^\intercal W)$. Unfortunately, not all of the Kronecker product's convenient algebraic properties carry over to the symmetric Kronecker product. For example, $(W \mathbin{\overline{\otimes}} W)^{-1} = W^{-1} \mathbin{\overline{\otimes}} W^{-1}$, but $(A \mathbin{\overline{\otimes}} B)^{-1} \neq A^{-1} \mathbin{\overline{\otimes}} B^{-1}$ in general, and inversion of this general form is straightforward only for commuting, symmetric $A, B$ [1]. This is why the proof for the following theorem is considerably more tedious than the one for Lemma 2.1.

THEOREM 2.3 (proof in Appendix A.2). *Assume a Gaussian prior of mean* $B_0$ *and covariance* $V = W \mathbin{\overline{\otimes}} W$ *on the elements of a symmetric matrix* $B$. *After* $M$ *linearly independent noise-free observations of the form* $Y = BS$, $Y, S \in \mathbb{R}^{N\times M}$, $\mathrm{rk}(S) = M$, *the posterior belief over* $B$ *is a Gaussian with mean*

$$\begin{aligned} B_M = B_0 &+ (Y - B_0S)(S^\intercal WS)^{-1}S^\intercal W + WS(S^\intercal WS)^{-1}(Y - B_0S)^\intercal \\ &- WS(S^\intercal WS)^{-1}(S^\intercal(Y - B_0S))(S^\intercal WS)^{-1}S^\intercal W, \end{aligned} \tag{2.8}$$

*and posterior covariance*

$$V_M = (W - WS(S^\intercal WS)^{-1}S^\intercal W) \mathbin{\overline{\otimes}} (W - WS(S^\intercal WS)^{-1}S^\intercal W). \tag{2.9}$$

This immediately leads to the following corollary.

COROLLARY 2.4. *The Dennis family of quasi-Newton methods is the posterior mean after one step* ($M = 1$) *of Gaussian regression on matrix elements.*

*Proof.* Assume $Y, S \in \mathbb{R}^{N\times 1}$, and set $c = WS$ in (1.2). ∎

Note that, for each member of the Dennis class, there is an entire vector space of $W$ consistent with $c = WS$. Additionally, each member of the Dennis family is itself a scalar space of choices $c$, because (1.2) is unchanged under the transformation $c \to \alpha c$ for $\alpha \in \mathbb{R}_{\setminus 0}$. Dealing with these degrees of freedom turns out to be the central task when defining probabilistic interpretations of linear solvers.

**2.2.2. Remark on the structure of the prior covariance.** The fact that symmetric Kronecker product covariance matrices give rise to some of the most popular secant methods may be reason enough to be interested in these structured Gaussian priors. This section provides two additional arguments in their favor.

The first argument, applicable to the entire family of Gaussian inference rules, is that they give consistent estimates and thus convergent solvers: The priors of Lemma 2.1 and Theorem 2.3 assign nonzero mass to all square and all symmetric matrices, respectively. It thus follows, from standard theorems about the consistency of parametric Bayesian priors (e.g., [6]), that linear solvers based on the mean estimate arising from either of these two Gaussian priors, applied to linear problems of general or symmetric structure, respectively, are guaranteed (assuming perfect arithmetic precision) to converge to the correct $B$ (and $B^{-1}$, where it exists) after $M = N$ linearly independent line searches (i.e., $\mathrm{rk}(S) = M$). This is because the Schur complement $W - WS(S^\intercal WS)^{-1}S^\intercal W$ is of rank $N - M$ [41, eq. 0.9.2], so the remaining belief after $M = N$ is a point-mass at the unique $B = YS^{-1}$. By a generalization of the same argument, it also follows that these linear solvers are always exact within the vector space spanned by the line-search directions. This holds for all choices of prior parameters $B_0$ and $W$, as long as $W$ is strictly positive definite. Of course, good convergence *rates* do depend crucially on these two choices. And the aim in this paper is to also identify choices for these parameters such that the posterior uncertainty around the mean estimate is meaningful, too.

Since we know $B$ to be positive definite, it would be desirable to restrict the prior explicitly to the positive definite cone. Unfortunately, this is not straightforward within the Gaussian family, because normal distributions have full support. A seemingly more natural prior over this cone is the Wishart distribution popular in statistics,

$$\mathcal{W}(B; W, \nu) \propto |B|^{\nu/2-(N-1)/2} \exp\left(-\frac{\nu}{2}\operatorname{tr}(W^{-1}B)\right) \tag{2.10}$$

(the $\propto$ symbol suppresses an irrelevant normalization constant). Using this prior in conjunction with linear observations, however, causes various complications, because the Wishart is not conjugate to one-sided linear observations of the form discussed above. So one may be interested in finding a "linearization" (a Gaussian approximation of some form) for the Wishart, for example through moment matching. And, indeed, the second moment (covariance) of the Wishart is $\nu^{-1}(W \otimes W)$ (see, e.g., [31]).

**3. Choice of parameters.** Having motivated the Gaussian hypothesis class, the next step is to identify individual desirable parameter choices in this class. The following Corollary follows directly from Theorem 2.3, by comparing (2.5) with (1.3) to (1.7). In each of the following cases, $\alpha \in \mathbb{R}_{\setminus 0}$.

Corollary 3.1.

1. *The Powell symmetric Broyden update rule is the one-step posterior mean for a Gaussian regression model with* $W = \alpha \boldsymbol{I}$.
2. *The Symmetric Rank-*1 *rule is the one-step posterior mean for a Gaussian regression model with the implicit choice* $W = \alpha(B - B_0)$. *(For a specific rank-*1 *observation, there is a linear subspace of choices* $W$ *which give* $WS = Y$, *but* $W = B$ *is the only globally consistent such choice).*
3. *The Greenstadt update rule is the one-step posterior mean for a Gaussian regression model with* $W = \alpha B_0$.
4. *The DFP update is the one-step posterior mean for the implicit choice* $W = \alpha B$. *(This choice is unique in a manner analogous to the above for SR*1*.)*
5. *The BFGS rule is the one-step posterior mean for the implicit choice* $W = \alpha\big(B + \sqrt{\frac{s^\intercal Bs}{s^\intercal B_0 s}}B_t\big)$. *(This, too, is unique in a manner analogous to the above.)*

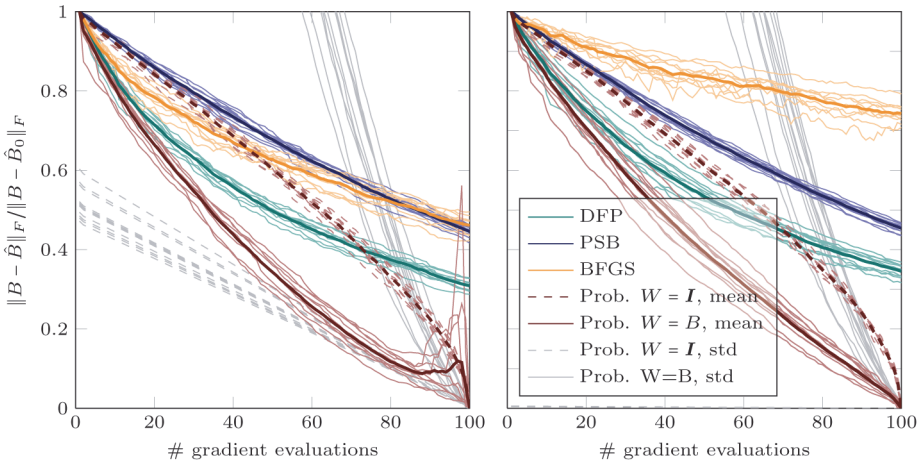


FIG. 1. *Effect of parameter choice and exact versus independent updates. Left: ten randomly generated linear problems with $N = 100$ with eigenvalue scale $\lambda = 10$. Right: Analogous problems with eigenvalue scale $\lambda = 1000$. Individual experiments as thin lines, means over all ten experiments as thick lines. The spikes for the $W = B$ estimate at the end of the left plot are numerical artifacts caused by ill-conditioned random projections. They do not arise in the optimization setting.*

It may seem circular for an inference algorithm trying to infer the matrix $B$ to use that very matrix as part of its computations (SR1, DFP, BFGS). But computation of the mean in (2.8) only requires the projections $BS$ of $B$, which are accessible because $BS = Y$. However, the posterior uncertainty (see (2.9)), which is not part of the optimizers in their contemporary form, cannot be computed this way.

Hence, with the exception of PSB, the popular secant rules all involve what would be called *empirical Bayesian* estimation in statistics, i.e., parameter adaptation from observed data. We also note again that the connection between probabilistic maximum-a-posterior estimates and Dennis-class updates only applies in the first of $M$ steps. As such, the Dennis updates ignore the dependence between information collected in older and newer search directions that leads to the matrix inverse of $G = (S^\intercal W S)$ in (2.8) and (2.9) (obviously, including this information explicitly requires solving $M$ linear problems, at additional cost). As will be shown in Lemma 3.3, though, for *some* members of the Dennis family, and for their use within *linear* problems, this simplification is, in fact, *exact*.

**3.1. A motivating experiment.** How relevant is the difference between the full rank-$2M$ posterior update and a sequence of $M$ rank-2 updates? Figure 1 shows results from a simple conceptual experiment. For this test only, the various estimation rules are treated as "stand-alone" inference algorithms, not as optimizers. Random positive definite matrices $B \in \mathbb{R}^{N\times N}$ were generated as follows: Eigenvalues $d_i, i = 1, \ldots, N$ were drawn i.i.d. from an exponential distribution $p(d) = {}^1\!/\!_\lambda \exp(-d/\lambda)$ with scale $\lambda = 10$ (small eigenvalues, left plot) or $\lambda = 1000$ (large eigenvalues, right plot), respectively. A random rotation matrix $Q \in SO(N)$ was drawn uniformly from the Haar measure over $SO(N)$, using the subgroup algorithm of Diaconis and Shahshahani [15], giving $B = QDQ^\intercal$ (where $D = \mathrm{diag}(d)$). Projections—simulated "search directions"—were drawn uniformly at random as $S \in \mathbb{R}^{N\times M}, s_{nm} \sim \mathcal{N}(0, 1)$. For $M = 1, \ldots, N$, the Powell Symmetric Broyden (PSB), DFP, and BFGS, as well the corresponding posterior means from (2.8) with $W = \boldsymbol{I}$ (equal to PSB after one step)

and $W = B$ (equal to DFP after one step) were used to construct point estimates $B_M$ for $B$. The plot shows the Frobenius norm $\|B_M - B\|_F$ between true and estimated $B$, normalized by the initial error $\|B_0 - B\|_F$. All algorithms used $B_0 = \boldsymbol{I}$.

Because directions $s$ were chosen randomly, these results say little about these algorithms as optimizers. What they do offer is an intuition for the difference between the exact rank-$2M$ posterior and repeated application of rank-2 Dennis-class update rules. A first observation is that, in this setup, keeping track of the dependence between consecutive search directions through $S^\intercal WS$ makes a big difference: For both pairs of "related" algorithms PSB and $W = \boldsymbol{I}$, as well as DFP and $W = B$, the full posterior mean dominates the simpler "independent" update rule. In fact, the classic secant rules do not converge to the true Hessian $B$ in this setup. The consistency argument in section 2.2.2 only applies to estimators constructed by exact inference. The experiment shows how crucial tracking the full Gram matrix $S^\intercal WS$ is after $M > 1$.

A second, not surprising, observation is that although both probabilistic algorithms are consistent—they converge to the correct $B$ after $N$ steps—the quality of the inferred point estimate after $M < N$ steps depends on the choice of parameters. The simpler $W = \boldsymbol{I}$ (PSB) choice performs qualitatively worse than the $W = B$ (DFP) choice.

The posterior covariances were used to compute posterior uncertainty estimates for $\|B_M - B\|_F$ (gray lines in Figure 1): The Frobenius norm can be written as $\|B_M - B\|_F^2 = \overrightarrow{(B_M - B)}^\intercal \overrightarrow{(B_M - B)}$; thus the expected value of this quadratic form is

$$\mathbb{E}[\overrightarrow{(B_M - B)}^\intercal \overrightarrow{(B_M - B)}] = \sum_{ij} V_{M,(ij)(ij)} = \sum_{ij} \frac{1}{2}(W_{M,ii}W_{M,jj} + W_{M,ij}W_{M,ij}), \tag{3.1}$$

with $W_M := W - WS(S^\intercal WS)^{-1}S^\intercal W$. (To be clear, for $W = B$, computing this uncertainty required the unrealistic step of giving the algorithm access to $B$, which only makes sense for this conceptual experiment.) The uncertainty estimate for $W = \boldsymbol{I}$ (dashed gray lines) is all but invisible in the right-hand plot because its values are very close to 0—this algorithm has a *badly calibrated uncertainty measure* that has no practical use as an estimate of error. The uncertainty under $W = B$ (solid gray lines), on the other hand, scales qualitatively with the size of $B$. This is because scaling $B$ by a scalar factor automatically also scales the covariance by the same factor. This has been noted before as a "nondimensional" property of BFGS/DFP [35, eq. 6.11]. However, it is also apparent that the uncertainty estimate is too large in both plots—here by about a factor of 5. To understand why, we consider the individual terms in the sum of (3.1) at the beginning of the inference: The ratio between the true estimation error on element $B_{ij}$ and the estimated error is

$$e_{ij}^2 = \frac{(B_0 - B)_{ij}^2}{\mathbb{E}[(B_0 - B)_{ij}^2]} = 2\frac{B_{ij}^2 - 2B_{ij}B_{0,ij} + B_{0,ij}^2}{W_{ii}W_{jj} + W_{ij}^2}. \tag{3.2}$$

One may argue that a "well-calibrated" algorithm should achieve $e_{ij} \approx 1$. A problem with the choice $W = B$ becomes apparent considering diagonal elements and $B_0 = \boldsymbol{I}$:

$$e_{ii}^2 = \frac{(B_{ii} - 1)^2}{B_{ii}^2} = \left(1 - \frac{1}{B_{ii}}\right)^2. \tag{3.3}$$

This means the DFP prior is well-calibrated only for large diagonal elements ($B_{ii} \gg 1$). For diagonal elements $B_{ii} \approx 1$, it is underconfident ($e_{ii} \nrightarrow 0$, estimating too large an error), and for very small diagonal elements $B_{ii} > 0, B_{ii} \ll 1$, it can be severely overconfident ($e_{ii} \nrightarrow \infty$ estimating too small an error). For off-diagonal elements and unit prior mean, the error estimate is

$$e_{ij}^2 = \frac{2B_{ij}^2}{B_{ij}^2 + B_{ii}B_{jj}} = \frac{2}{1 + B_{ii}B_{jj}/B_{ij}^2} \qquad \text{for } i \neq j. \tag{3.4}$$

For positive definite $B$, $e_{ij}^2 < 1$ off the diagonal holds because, for such matrices, $B_{ij}^2 < B_{ii}B_{jj}$ (see, e.g., [28, Corollary 7.1.5]), but of course $e_{ij}^2$ can still be very small or even vanish, e.g., for diagonal matrices. It is possible to at least fix the overconfidence problem, using the degree of freedom in Corollary 3.1 to scale the prior covariance to $W = \theta^2 B$ with $\theta = \lambda_{\min}/(\lambda_{\min} - 1)$, using $\lambda_{\min}$, the smallest eigenvalue of $B$. This at least ensures $e_{ij} \leq 1 \; \forall (i,j)$.

Interestingly, setting $W = B - B_0$ (which gives the SR1 rule after the first observation, but not after subsequent ones) gives $e_{ii}^2 = 1$, and $e_{ij} < 1$ for $i \neq j$. It also has the property that the norm of the true $B$ under this prior is

$$\overrightarrow{(B - B_0)}^\intercal ((B - B_0) \otimes (B - B_0))^{-1} \overrightarrow{(B - B_0)} = \overrightarrow{\boldsymbol{I}}^\intercal \overrightarrow{\boldsymbol{I}} = N, \tag{3.5}$$

so the true $B$ is exactly one standard deviation away from the mean under this prior. These properties suggest this covariance, which will be called *standardized norm* covariance, for further investigation in section 4, which addresses the question: Is it possible to construct a linear solver that, without "cheating" (using $B$ or $H$ explicitly in the covariance), has a well-calibrated uncertainty measure and can thus meaningfully estimate the error of its computation; ideally, without major cost increase?

**3.2. Structure of the Gram matrix.** The above established that, treated as inference rules for matrices, general Dennis rules are probabilistically exact only after one rank-1 observation $y = Bs$. How strong is the error thus introduced? In fact, as the following lemma shows, there are choices of search directions $S$ for which the existing algorithms do become exact probabilistic inference.

LEMMA 3.2 (proof in Appendix A.3). *If the Gram matrix $S^\intercal W S$ is a diagonal matrix (i.e., if the search directions $S \in \mathbb{R}^{N \times M}$ are conjugate under the covariance parameter $W$), then the $M$ repeated rank-2 update steps of classic secant-rule implementations result in an estimate that is equal to the posterior mean under exact probabilistic Gaussian inference from $(Y, S)$. (The equivalent statement for inverse updates requires conjugacy of the $Y$ under $W$.)*

So a *cheap*[4] probabilistic optimizer can be constructed by choosing search directions conjugate under $W$. The following reformulation of a previously known lemma[5] shows that, in fact, both the BFGS and DFP update rules have this property.

LEMMA 3.3 (additional proof in Appendix A.4). *For linear problems $Bx = b$ with symmetric positive definite $B$ and exact line searches, under the DFP covariance $W = B$, and line searches along the inverse of the posterior mean of the Gaussian belief, the*

[4]We note in passing that, to reduce cost further, and regardless of whether the Gram matrix is diagonal or not, the updates of (2.5) can be *approximated* by using only the $\tilde{M}$ most recent pairs $(s_i, y_i)$, or by retaining a restricted rank $\tilde{M}$ form of the update. This is the analogue to "limited-memory" methods [34] well-known in the literature for large-scale problems.

[5]This result is quoted by Nazareth in 1979 [33] as "well-known," with a citation to [32]. The proof in the appendix is less general, but may help put this lemma in the context of this text.

*Gram matrix is diagonal. Analogously for inverse updates: For inference on $H = B^{-1}$ under the BFGS covariance $W = H$ on the same linear optimization problem and line searches along the posterior mean over $H$, the Gram matrix is diagonal.*

The following result by Nazareth [33] establishes that, for linear problems, the inference interpretation for BFGS transfers directly to the conjugate gradient (CG) method of Hestenes and Stiefel [27].

THEOREM 3.4 (Nazareth [33][6]). *For linear optimization problems as defined in Lemma 3.3, BFGS inference on $H$ with scalar prior mean, $H_0 = \alpha \boldsymbol{I}, \alpha \in \mathbb{R}$, is equivalent to the conjugate gradient algorithm in the sense that the sequence of search directions is equal: $s_i^{BFGS} = s_i^{CG}$.*

The connection between BFGS and CG is intuitive within the probabilistic framework: BFGS uses $W = H$, so its mean estimate $H_M$ is the "best guess" for $H$ under (the minimizer of) the norm $\overrightarrow{(H - H_M)}^\intercal (H \otimes H)^{-1} \overrightarrow{(H - H_M)}$, and its iterated estimate $x_M$ is the best rank-$M$ estimate for $x$ when the error is measured as $(x - x_M)^\intercal W^{-1}(x - x_M) = (x - x_M)^\intercal B(x - x_M)$. Minimizing this quantity after $M$ steps is a well-known characterization of CG [35, eq. 5.27].

Theorem 3.4 implies that describing BFGS in terms of Gaussian inference also gives a Gaussian interpretation for CG "for free." From the probabilistic perspective, and exclusively for linear problems, CG is "just" a compact implementation of iterated Gaussian inference on $H$ from $p(H) = \mathcal{N}(H; \boldsymbol{I}, H \otimes H)$, with search directions along $H_M F(x_M) = H_M(Bx_M - b)$. This observation has conceptual value in itself (the natural question, left open here, is what it implies for the nonparametric extensions of CG). But Theorem 3.4, among other things, also implies the following helpful properties for the search directions $s_i$ chosen by, and gradients $F_i$ "observed" by, the (scalar prior mean) BFGS algorithm. They are all well-known properties of the conjugate gradient method (e.g., [35, Thm. 5.3]). In the following, generally assume that the algorithm has not converged at step $M < N$, and remember that the $F_M = Bx_M - b$ are the residuals (gradients of $f(x) = 1/2 x^\intercal Bx - x^\intercal x$) after $M$ steps, which form $y_M = F_M - F_{M-1}$. We have that

- the set of evaluated gradients / residuals is orthogonal:

$$F_i^\intercal F_j = 0 \text{ for } i \neq j \text{ and } i, j < N; \tag{3.6}$$

- the gradients (and thus $Y$) span the Krylov subspaces generated by $(B, b)$:

$$\operatorname{span}\{F_0, F_1, \ldots, F_M\} = \operatorname{span}\{F_0, BF_0, \ldots, B^M F_0\}; \tag{3.7}$$

- line searches and gradients span the same vector space:

$$\operatorname{span}\{s_0, s_1, \ldots, s_M\} = \operatorname{span}\{F_0, BF_0, \ldots, B^M F_0\}. \tag{3.8}$$

**3.3. Discussion.** We have established a probabilistic interpretation of the Dennis class of quasi-Newton methods, and the CG algorithm, as Gaussian inference: The Dennis class can be seen as Gaussian posterior means after the first line search (Corollary 2.4), but this connection extends to multiple search directions only if the search directions are conjugate under prior covariance (Lemma 3.2). For linear problems, this is the case for the DFP, BFGS update rules (Lemma 3.3). Since BFGS is equivalent to CG on linear problems (Lemma 3.4), this also establishes a probabilistic

[6] Dixon [16, 17] provided a related result linking CG to the whole Broyden class of quasi-Newton methods: they become equivalent to CG when the starting matrix is chosen as the preconditioner.

interpretation for linear CG. These results offer new ways of thinking about linear solvers, in terms of solving an inference problem by collecting information and building a model rather than by designing a dynamic process converging to the minimum of a function. It is intriguing that, from this vantage point, the extremely popular CG/BFGS methods look less well-calibrated than one may have expected (section 3.1).

The obvious next question is: Can one design explicitly uncertain linear solvers with a reasonably well-calibrated posterior? In addition to the scaling issues, a challenge is that, for BFGS/CG, the prior covariance $W = H$ is only an implicit object. After $M < N$ steps, there exists a $1/2(N-M)(N-M+1)$-dimensional cone of positive definite covariance matrices fulfilling $WY = S$ (and, additionally, a scalar degree of freedom inherent to the Dennis class). How do we pick a point in this space?

**4. Constructing explicit posteriors.** The remainder of this paper will focus exclusively on inference on $H = B^{-1}$, on inverse update rules, priors $p(H) = \mathcal{N}(\overrightarrow{H}; \overrightarrow{H}_0, W \otimes W)$. As pointed out in section 1.2, these arise from the direct rules under exchange of $S$ and $Y$: Given $(S, Y) \in \mathbb{R}^{N\times M}$, the posterior belief is $p(H \mid S, Y) = \mathcal{N}(\overrightarrow{H}; \overrightarrow{H}_M, W_M \otimes W_M)$ with

$$\begin{aligned} H_M &= H_0 + (S - H_0Y)(Y^\intercal WY)^{-1}Y^\intercal W + WY(Y^\intercal WY)^{-1}(S - H_0Y)^\intercal \\ &\quad - WY(Y^\intercal WY)^{-1}[Y^\intercal(S - H_0Y)](Y^\intercal WY)^{-1}Y^\intercal W, \\ W_M &= W - WY(Y^\intercal WY)^{-1}Y^\intercal W. \end{aligned} \tag{4.1}$$

Recall from sections 1.2 and Corollary 3.1 that, cast as an inverse update, BFGS (CG) arises from the prior $p(H) = \mathcal{N}(H; \boldsymbol{I}, \theta^2(H \otimes H))$ for arbitrary $\theta \in \mathbb{R}_+$.

**4.1. Fitting covariance matrices.** Equations (1.6), (1.7) show that both the BFGS and DFP priors in principle require access to $H$. As noted above, for this mean estimate it is implicitly feasible to use $W = H$, because this computation only requires observed projections $WY = HY = S$. Computing the covariance under $W = H$, however, can only be an idealistic goal: After $M$ steps, only a subspace of rank $\frac{1}{2}(N(N+1) - (N-M)(N-M+1))$ of the elements of $H$ is identified. To see this explicitly, consider the singular value decomposition[7] $Y = Q\Sigma U^\intercal$, which defines a symmetric positive definite $T \in \mathbb{R}^{N\times N}$ through $W = QTQ^\intercal$. This notation gives

$$W_M = W - WY(Y^\intercal WY)^{-1}Y^\intercal W = Q(T - T\Sigma(\Sigma^\intercal T\Sigma)^{-1}\Sigma^\intercal T)Q^\intercal. \tag{4.2}$$

Considering the structure of $\Sigma$, one can write $T$ in terms of block matrices

$$T = \begin{pmatrix} T_{++} & T_{+-} \\ T_{-+} & T_{--} \end{pmatrix} \qquad \text{then} \qquad W_M = Q \begin{pmatrix} 0 & 0 \\ 0 & T_{--} - T_{-+}T_{++}^{-1}T_{+-} \end{pmatrix} Q^\intercal, \tag{4.3}$$

with $T_{++} \in \mathbb{R}^{M\times M}, T_{-+} = T_{+-}^\intercal \in \mathbb{R}^{M\times(N-M)}, T_{--} \in \mathbb{R}^{(N-M)\times(N-M)}$ (and positive definite $T_{++}$, a principal block of the positive definite $T$). Observing $(S, Y)$ exactly identifies $[T_{++}, T_{+-}]^\intercal = QSU^\intercal D^{-1}$, and provides no information at all[8] about $T_{--}$.

[7] This is using the convention that the orthonomal matrices are $Q \in \mathbb{R}^{N\times N}$ and $U \in \mathbb{R}^{M\times M}$, and $\Sigma \in \mathbb{R}^{N\times M}$ can be written with orthonormal $Q \in \mathbb{R}^{N\times N}$ and $U \in \mathbb{R}^{M\times M}$, and rectangular diagonal $\Sigma \in \mathbb{R}^{N\times M}$, which can be written as $\Sigma = [D, \mathbf{0}]^\intercal$ with an invertible diagonal matrix $D \in \mathbb{R}^{M\times M}$ and empty $\mathbf{0} \in \mathbb{R}^{M\times(N-M)}$.

[8] Knowing $H$ to be positive definite does provide a lower bound on the eigenvalues of $T_{--}$.

A primary goal in designing a probabilistic linear solver is thus, at step $M < N$, to (1) identify the span of $W_M$, ideally without incurring additional cost, and to (2) fix the entries in the remaining free dimensions in $W_M$, by using some regularity assumptions[9] about $H$. The equivalence between BFGS and CG offers an elegant way of solving problem (1) with no additional computational cost: Recall from Theorem 2.3 and Lemma 3.2 that the covariance after $M$ steps under $W = H$ is $W_M \otimes W_M$ with

$$W_M = W - \sum_i^M \frac{W y_i (W y_i)^\intercal}{s_i^\intercal y_i} = W - \sum_i^M \frac{s_i s_i^\intercal}{s_i^\intercal y_i} = W - S(S^\intercal Y)^{-1} S^\intercal. \tag{4.4}$$

Because, by (3.8) the vector-space spanned by $S$ is identical to that spanned by the *orthogonal* gradients, we can write the space of all symmetric positive semidefinite matrices $W$ with the property $WY = S$ as

$$W(\Omega) = S(S^\intercal Y)^{-1} S^\intercal + (\boldsymbol{I} - \bar{F}\bar{F}^\intercal)\Omega(\boldsymbol{I} - \bar{F}\bar{F}^\intercal), \tag{4.5}$$

with the right-orthonormal matrix $\bar{F}$ containing the $M$ normalized gradients $F_i/\|F_i\|$ in its columns, and a positive definite matrix $\Omega \in \mathbb{R}^{N\times N}$ (the effective size of the space spanned in this way is only $\mathbb{R}^{(N-M)\times(N-M)}$, so $\Omega$ is overparameterizing this space).

**4.1.1. Standardized norm posteriors using conjugate gradient observations.** Equation (4.5) parametrizes posterior covariances of the BFGS family. In light of the scaling issues of these priors discussed in section 3.1, one would prefer, from the probabilistic standpoint, to use the standardized norm priors $p(H) = \mathcal{N}(H; \alpha\boldsymbol{I}, (H - H_0)\otimes(H - H_0))$, but these priors do not share BFGS/CG's other good numerical properties. Instead, a hybrid algorithm can be constructed as follows:

1. Solve the linear problem using the conjugate gradient method. While the algorithm runs, collect $S, Y, \bar{F}$. This has a storage cost of $2NM + M$ floats: Because $Y$ consists of differences between subsequent columns of $F$, it does not need to be stored explicitly; the column norms $\|F\|_i$ required to compute $\bar{F}$ require $M$ extra floats. The computation cost of the standard conjugate gradient algorithm is $\mathcal{O}(M)$ matrix-vector multiplications (that is, $\mathcal{O}(MN^2)$ assuming a dense matrix), plus $\mathcal{O}(MN)$ operations for the algorithm itself (including computation of $\|F\|_i$).
2. *Using the* $(S, Y, \bar{F})$ *constructed by CG*, compute the *standardized-norm* posterior on $H$, i.e., use the prior $p(H)$ defined above, which yields a Gaussian posterior with mean and covariance

$$H_M = H_0 + (S - H_0 Y)(Y^\intercal (S - H_0 Y))^{-1}(S - H_0 Y)^\intercal \tag{4.6}$$

$$= \alpha\boldsymbol{I} - (S - \alpha Y)(Y^\intercal S - \alpha Y^\intercal Y)^{-1}(S - \alpha Y)^\intercal \qquad \text{and} \tag{4.7}$$

$$W_M = (H - H_0) - (S - H_0 Y)(Y^\intercal (S - H_0 Y))^{-1}(S - H_0 Y)^\intercal \tag{4.8}$$

$$= S(S^\intercal Y)^{-1} S^\intercal + (\boldsymbol{I} - \bar{F}\bar{F}^\intercal)\Omega(\boldsymbol{I} - \bar{F}\bar{F}^\intercal) - \alpha\boldsymbol{I} - (S - \alpha Y)(Y^\intercal S - \alpha Y^\intercal Y))^{-1}(S - \alpha Y)^\intercal. \tag{4.9}$$

[9] A probabilistically more appealing approach would be to use a hyper-prior on the elements of $W$, marginalized over the unidentified degrees of freedom. It is currently unclear to the author how to do this in a computationally efficient way.

A prerequisite for this is to choose $\alpha < \lambda_{\min}(H)$, less than the smallest eigenvalue of $H$, to ensure that $W = H - H_0$ is positive definite. But $\lambda_{\min}(H) = 1./\lambda_{\max}(B)$, which can be estimated efficiently (and without additional cost) from the $\|F\|_i$. Another minor hurdle is that (4.7) and (4.9) require the inverse of $S^\intercal Y - \alpha Y^\intercal Y$. The columns of $Y$ are $Y_i = F_i - F_{i-1}$, so, because conjugate gradient constructs orthogonal gradients, $Y^\intercal Y$ is a symmetric tridiagonal matrix, $Y_i^\intercal Y_j = \delta_{ij}(\|F_i\|^2 + \|F_{i-1}\|^2) + (\delta_{i(j-1)} + \delta_{(i+1)j})\|F_i\|^2$, and $S^\intercal Y$ is diagonal because the $S$ are conjugate under $B$. So the entire Gram matrix is tridiagonal, and the $M$ linear problems in $(Y^\intercal S - \alpha Y Y^\intercal)^{-1}(S - \alpha Y)^\intercal$ can be solved in $\mathcal{O}(M^2)$, e.g., using the Thomas algorithm [7, Alg. 4.3].

3. Using some statistical estimation rule to be defined below, estimate $\Omega$. Section 4.2 proposes several such rules of cost $\mathcal{O}(M)$, but as usual for statistical estimation problems, there is no uniquely correct way to perform this step.

While there is a vague connection between the standardized norm prior and the SR1 algorithm by Corollary 3.1, the algorithm described above is quite different from the SR1 method. It uses search directions constructed by BFGS/CG, and its update rule uses the exact Gram matrix, not the repeated rank-1 updates that give SR1 its name.

*Computational cost.* The computation overhead of constructing this posterior mean and covariance, after running the conjugate gradient algorithm, is $\mathcal{O}(M^2)$, which is small even compared to the internal $\mathcal{O}(MN)$ cost of CG, let alone the $\mathcal{O}(MN^2)$ for the matrix-vector multiplications in CG. Storing the posterior mean and covariance requires $\mathcal{O}(NM)$ space, which is feasible even for relatively large problems. Crucially, retaining the covariance adds almost no overhead to storing the mean alone.

**4.2. Estimation rules.** The remaining step is to find estimates for $\Omega$. It is clear that there are myriad options for fixing such rules. For an initial evaluation, we adopt the perhaps simplistic but straightforward approach of estimating $\Omega$ to a scalar matrix $\Omega = \omega^2 \boldsymbol{I}$ (one way to motivate this is to argue that, at step $M$, future line searches $s_{M+i}$ will point in an unknown direction in the span of $\boldsymbol{I} - \bar{F}\bar{F}^\intercal$, so it makes sense not to prefer any direction in the choice of $\Omega$).

A natural idea is to use regularity structure on quantities already computed during the run of the conjugate gradient algorithm: Assume the algorithm is currently at step $T$. If at step $M < T$ we had tried to predict the Gram matrix diagonal element $y_{M+1}^\intercal W y_{M+1} = -s_{M+1}^\intercal F_M$ using the structure for $W$ described above, we would have predicted, because $F_M$ is known to be in the span of $S$, and orthogonal to $(\boldsymbol{I} - \bar{F}\bar{F}^\intercal)$,

$$y_{M+1}^\intercal W y_{M+1} = F_M^\intercal S (S^\intercal Y)^{-1} S^\intercal F_M + F_{M+1}^\intercal \Omega F_{M+1}, \tag{4.10}$$

$$-s_{M+1}^\intercal F_M = -\sum_{i=1}^{M} \frac{(F_M^\intercal s_i)^2}{s_i^\intercal F_{i-1}} + \omega^2 \|F_{M+1}\|^2, \tag{4.11}$$

$$\text{and thus} \qquad \omega^2 = \|F_{M+1}\|^{-2} \left[ \sum_{i=1}^{M} \frac{(F_M^\intercal s_i)^2}{s_i^\intercal F_{i-1}} - s_{M+1}^\intercal F_M \right]. \tag{4.12}$$

$\|F_{M+1}\|$ can be estimated from the norm of preceding gradients. The second term on the right-hand side of (4.12) is known at step $M$. The first term of the right-hand side can be estimated by regression, in ways further explored below.

First, to confirm that $\omega$ indeed tends to have regular structure related to the eigenvalue spectrum of $H$, Figure 2, right column, shows $\omega_i$ for $i = 1, \ldots, M$ during runs of CG on twenty linear problems, sampled from three different generative processes for $B = QDQ^\intercal \in \mathbb{R}^{200\times 200}$. In each case, orthonormal matrices were drawn uniformly

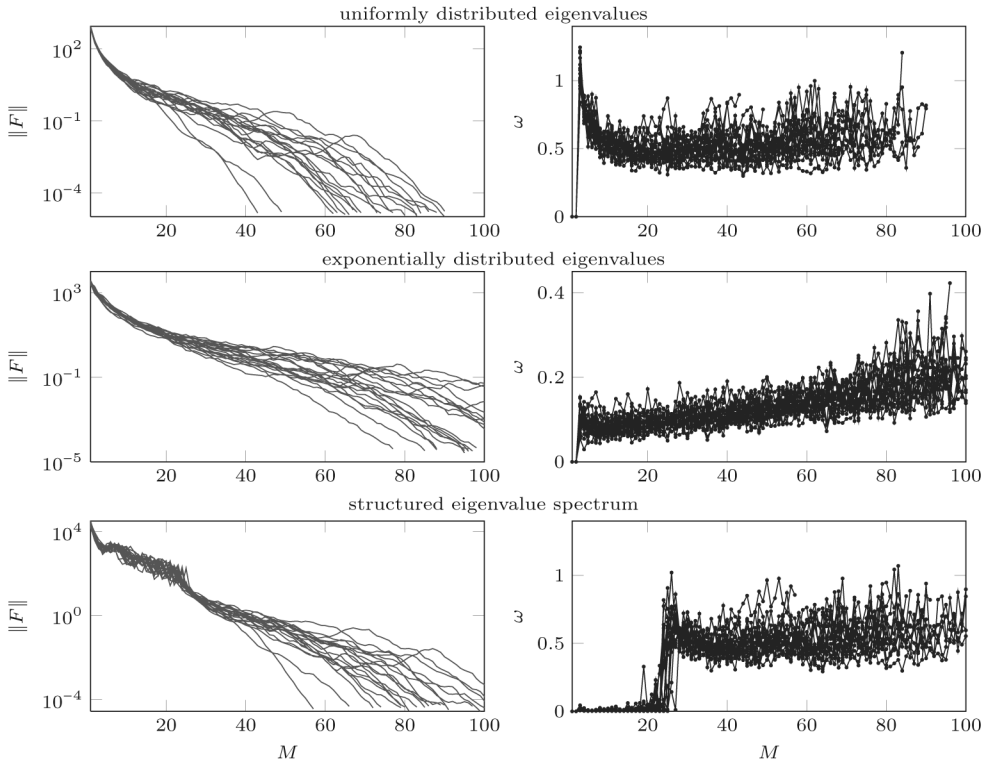


FIG. 2. *Fitting posterior uncertainty during iterative solution of linear problems for three different generative processes of $B$. Each plot shows results from twenty randomly generated experiments with, top row: uniformly; middle row: exponentially distributed eigenvalues; bottom row: structured eigenvalue spectrum (details in text). Left: Residual (gradient) $Bx - b$ as a function of number of line searches. Right: Projections $\omega = s_M^\intercal F_{M-1}$, whose regular structure is used for estimating $W(\Omega)$.*

from the Haar measure over $SO(N)$ as in section 3.1. For the top row of Figure 2, the eigenvalues (elements of $D = \mathrm{diag}(d)$) were drawn uniformly from $p(d_i) = U(0, 10)$ (the uniform distribution over $[0, 10]$). For the middle row, eigenvalues were drawn from the exponential distribution $p(d_i) = 1/\lambda \exp(-d_i/\lambda)$ with scale $\lambda = 10/\log 2$ (giving a median eigenvalue of 10). Finally, for the bottom row, eigenvalues were drawn from a structured process, with $d_i$ for $i = 1, \ldots, 20$ drawn from $p(d) = U(0, 10^3)$, and $d_i$ for $i = 21, \ldots, 200$ drawn from $p(d) = U(0, 10)$ (i.e., the corresponding eigenvalues of $H$ lie nonuniformly in $[0, 10^{-3}]$ and $[0, 0.1]$). Clear structure is visible in all cases. Using these observations, several different regression schemes for $\omega$ can be adopted.

- A simple baseline is a stationary model for the $\omega_i$. This was used to construct error estimates in Figures 3 to 5 (in gray for the middle and bottom row, black for the top row). Of course, if the eigenvalues of $B$ are uniformly distributed in the top row, the eigenvalues of $H$ (their inverses) are not.
- A slightly more elaborate model is a linear trend with noise: $\omega_i = ai + b + n$ (with $n \sim \mathcal{N}(0, \sigma^2)$). Linear regression on the values of $\omega_i$ can be performed in $\mathcal{O}(M)$. We can then set $\Omega = \bar{\omega}\boldsymbol{I}$ with $\bar{\omega} = aN + b$ the expected largest value of $\omega_i$ (i.e., a noisy upper bound). This approach was used to construct the (black) error estimates in the middle rows of Figures 3 to 5.

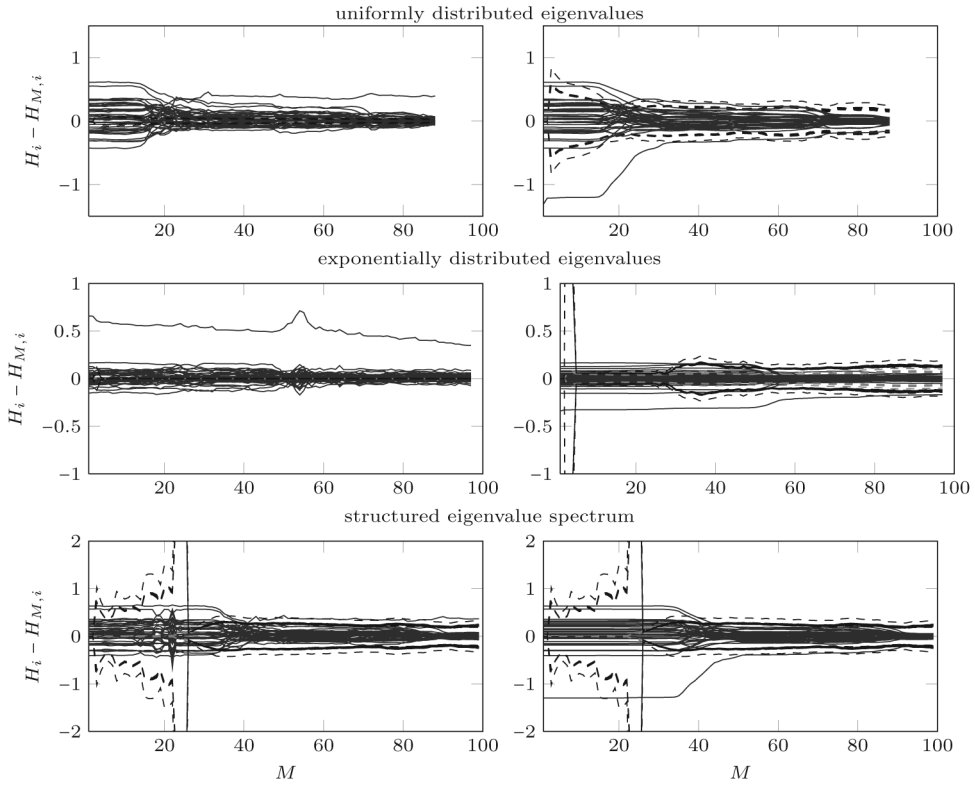


FIG. 3. *Error estimation on $H$. Posterior mean (solid red) and one standard deviation (dashed black, gray). Left: BFGS/CG prior. Right: Standardized norm prior, from CG observations. Rows as in Figure 2. The cut-off error bars in the bottom right plot rise up to values $< 6$.*

- Finally, if structural knowledge is available, e.g., that the first $L$ eigenvalues of $B$ are $\alpha$ times larger than the later ones, one may use the stationary rule from above, but explicitly multiply the estimate $\omega$ by $\alpha$ for the first $L$ steps. This may seem contrived, but, in fact, it is not uncommon in applications to know an effective number of degrees of freedom in $B$. For example, in nonparametric least-squares regression with a very large number of $N$ data points distributed approximately uniformly over a range of width $\rho$, using an RBF kernel of length scale $\lambda$, the model's number of degrees of freedom is $L = \rho/(2\pi\lambda)$ [38, eq. 4.3]. This rule was used to construct (black) error estimates in the bottom rows of Figures 3 to 5.

**4.3. Estimating quantities of interest.** This final part demonstrates a few example uses of the Gaussian posterior $p(H) = \mathcal{N}(\overrightarrow{H}; \overrightarrow{H}_M, W_M \otimes W_M)$ on $H$ constructed by the BFGS/CG method. Figures 3 to 5 show three such uses, explained below. Each row of this figure uses data from one of the experiments shown in the corresponding row of Figure 2.

**4.3.1. Estimating *H* itself.** The most obvious question is how far the estimate $H_M$ for $H$ after $M$ steps is from the true $H$. This distance is estimated directly by the Gaussian posterior of (4.1). The marginal distribution on any linear projection $A\overrightarrow{H}$ is $\mathcal{N}(A\overrightarrow{H}; A\overrightarrow{H}_M, A(W_M \otimes W_M)A^\intercal)$. In particular, the marginal distribution on

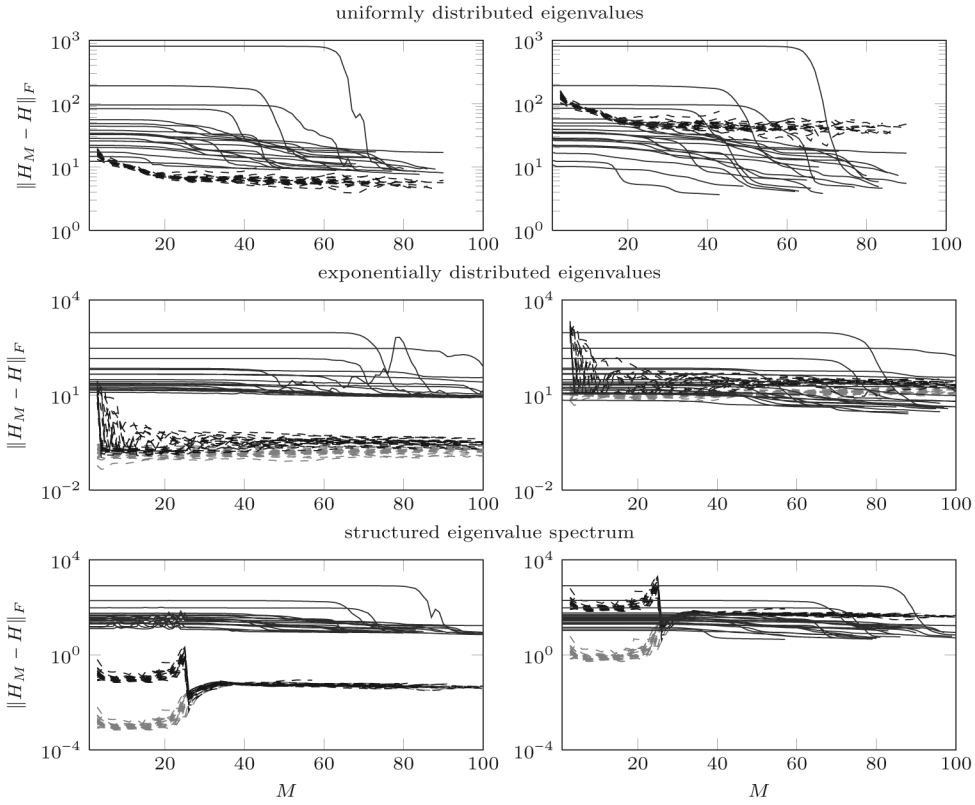


FIG. 4. *True and estimated norm error $\|H - H_M\|_F$. Posterior mean (solid red in electronic version) and one standard deviation (dashed black and dashed gray). Left: BFGS/CG prior. Right: Standardized norm prior, from CG observations. Rows as in Figure 2.*

each element $H_{ij}$ is a scalar Gaussian

$$p(H_{ij} \mid S_M, Y_M) = \mathcal{N}[H_{ij}; H_{M,ij}, 1/2(W_{M,ii}W_{M,jj} + W_{M,ij}^2)]. \tag{4.13}$$

Figure 3 shows this error estimate for forty elements of one particular $H$ (drawn uniformly at random from the $4 \cdot 10^4$ elements of the $200 \times 200$ matrix). The estimate arising from the uniform estimation rule for $\omega$ from section 4.2 is shown in gray in each panel (black for the top panel). The same quantity, estimated with the linear regression and structured estimation rules from section 4.2, is shown in black in the middle and bottom row, respectively. The left column of the figure shows results from the BFGS/CG prior; the right column shows results using the standardized norm prior on data constructed with the CG algorithm as described in section 4.1.1. As expected from the argument in section 3.1, the BFGS estimates are regularly considerably too small, while the standardized-norm estimates have a meaningful width. The error estimators have varying behavior. For the exponential eigenvalue spectrum, the estimator fluctuates strongly in the first few steps before settling to a good value (this could be corrected using a regularizer, left out here to not bias the results). For the structured-eigenvalues problems, the region around the step from small to large eigenvalues is problematic. But overall, they do provide a meaningful notion of error. In particular, they are rarely too small. For most uses of statistical error estimators,

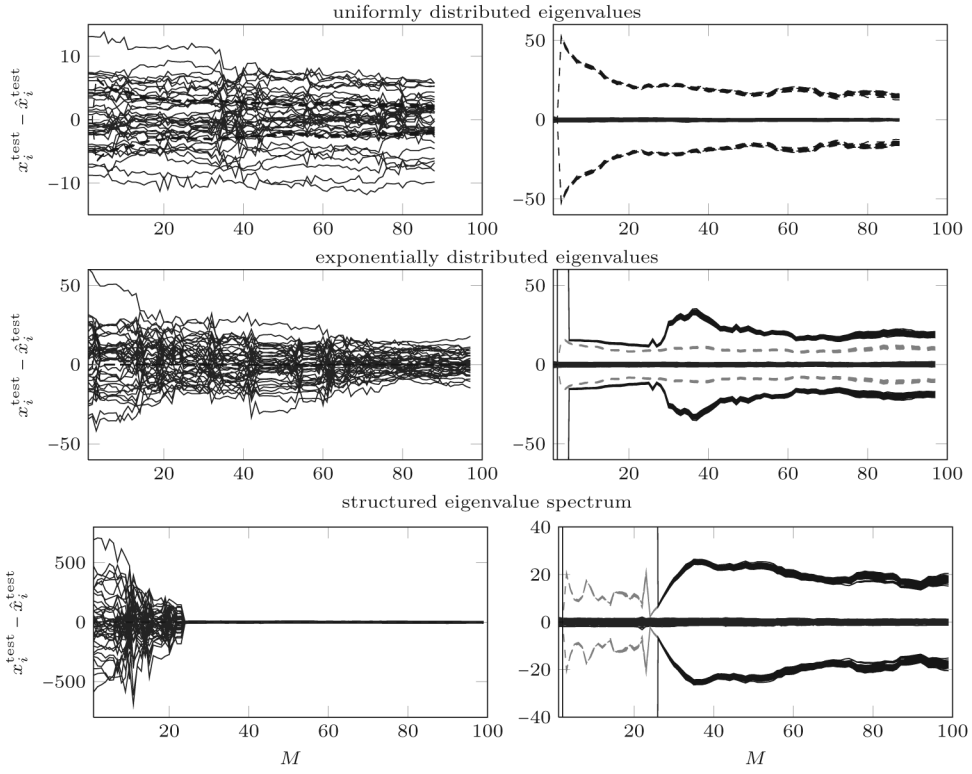


FIG. 5. *Estimating solutions to $Bx' = b'$. Elementwise error on a single test vector $x_{test}$. True error in blue (in electronic version). Error estimate with stationary model for $\omega$ in gray. Error estimate for model-specific estimate for $\omega$ (as in Figure 2) in black. Left: BFGS/CG prior. Right: Standardized norm prior, from CG observations. Rows as in Figure 2.*

it is better to be too conservative (too large) than to be too confident. Of course, it would be great if future research would find better calibrated error estimates.

As explained in (3.1), the same error estimates can also be collapsed into an error estimate on the norm $\|H - H_M\|_F$. Figure 4 shows results from such an experiment for the twenty different $H$'s from Figure 2. The quantitative results are similar to the previous figure, but this figure more clearly shows the difference between the baseline (gray) and exponential, structured error estimates (black), and the behavior of the estimated errors relative to the varying norms of the drawn $H$'s.

**4.3.2. Estimating solutions for new linear problems.** An obvious use for the estimate for $H$ found by CG/BFGS when solving *one* linear problem $Bx = b$ is as an instantaneous solution estimate for *other* linear problems $Bx_{\text{test}} = b_{\text{test}}$. The left and middle columns of Figure 5 show this use. In each case, an $x_{\text{test}}$ was drawn from $\mathcal{N}(x; 0, 10\boldsymbol{I})$, and the corresponding $b_{\text{test}} = Bx_{\text{test}}$ presented to the algorithm. Since $x_{\text{test}} = Hb_{\text{test}}$ is a linear projection of $H$, the posterior marginal on $x_{\text{test}}$ is also Gaussian $p(x_{\text{test}} \,|\, S_M, Y_M) = \mathcal{N}(x_{\text{test}}, H_M b_{\text{test}}, \Sigma)$, and has covariance matrix elements

$$\text{cov}(x_{\text{test},i}, x_{\text{test},j}) = \Sigma_{ij} = {}^1\!/_2(W_{ij} x_{\text{test}}^\intercal W x_{\text{test}} + (W x_{\text{test}})_i (W_{\text{test}})_j. \tag{4.14}$$

Figure 4 shows the true errors on the elements of $x_{\text{test}}$ (solid), and the estimated marginal errors (the diagonal elements of $\Sigma$) in dashed black for the stationary, linear, structured models, respectively (and, as in previous figures, the stationary model in dashed gray in the two nonstationary cases). More drastically than the previous ones, these figures show that the BFGS posterior can severely underestimate the error on elements of $x_{\text{test}}$, while the standardized norm prior at least provides outer bounds (albeit ones sometimes quite loose).

*Remark on convergence.* The error on $x_{\text{test}}$ does not always collapse over the course of finding $x$. This says more about CG as such than about its probabilistic interpretation: CG does not aim to construct $H$, but only to find $x_*$. For simplicity of exposition, we have assumed that $H = B^{-1}$ exists, and CG requires the full $N$ steps to converge, thus identifying $B$ and $H$. In general, CG regularly converges much earlier. For an intuition, consider the special case where $x_0 = 0$ and $b = [1, \dots, 1, 0, 0, \dots, 0]$ consists of $K$ consecutive ones and $N - K$ zeros. The CG/BFGS algorithm will never explore the lower $(N-K)\times(N-K)$ block of $H$, which may contain arbitrary numbers. If the primary aim is not $x_* = Hb$ but $H$ itself, a more elaborate course is needed; e.g., choosing several $b$ to span a space of interest over $H$. It is an interesting open question whether the probabilistic interpretation can be used to *actively* collapse the uncertainty on $H$ in a *typically* more efficient way than established matrix inversion methods like Gauss–Jordan (which is also a conjugate direction method [27]).

**5. Conclusion and outlook.** This text developed a probabilistic interpretation of iterative solvers for linear problems $Bx = b$ with symmetric $B$. The Dennis family of secant updates can be derived as the posterior mean of a parametric Gaussian model after one rank-1 observation. For rank $M$ observations, the match between these updates and Gaussian inference only holds if the search directions are conjugate under the prior covariance. This is the case for the DFP direct and BFGS inverse updates rules. Their equivalence to CG in the linear case makes them particularly interesting. However, it also became apparent that, from an inference perspective, the BFGS rule does not yield a well-scaled error measure.

As a first step toward a better scaled Gaussian posterior distribution, the standardized norm covariance was proposed. It is inspired by the SR1 rule, but leads to probabilistic corrections in the form of off-diagonal terms, and can be used with data produced by the CG algorithm, thus retaining the good numerical properties of that method. The space of possible covariance matrices consistent with the resulting mean is a subspace of the positive definite cone, which collapses during the run of the algorithm (the same holds for the BFGS/CG method). Several possible estimation rules for choosing elements in this space of covariances where proposed, arising from different structural assumptions over $H$. The resulting Gaussian posterior provides joint uncertainty estimates on the elements of $H$ and all linear projections of $H$, in particular of other linear problems $x_{\text{test}} = Hb_{\text{test}}$. This adds functionality to the conjugate gradient method, at a computational overhead much smaller than the cost of CG itself.

The implications for *non*linear optimization methods of both the quasi-Newton and CG families remain interesting open questions. For example, clearly the conjugacy assumption implicit in the Dennis class members is inconsistent with the probabilistic interpretation. This was already noted by Hennig and Kiefel [25, 26], who also proposed using a nonparametric Gaussian formulation to give a more explicit inference interpretation to nonlinear optimization. This left unanswered questions regarding the choice of prior covariance, which are only made more pressing by the results pre-

sented here. Another direction is inference from noisy evaluations, in which case the posterior covariance does not collapse to zero after finitely many steps of optimization, not even in the linear case. Some related results where previously discussed in [24], but the study of probabilistic numerical optimization remains at an early stage.

**Appendix. Proofs for results from main text.** Throughout the appendix, the notation $\Delta = Y - B_0 S$ will be used to represent the residual.

**A.1. Proof for Lemma 2.2.** Because the operator $\Gamma$ maps $\sum_{k\ell} \Gamma_{ij,k\ell} A_{k\ell} = \frac{1}{2}(A_{ij} + A_{ji})$ for all $A$, its elements can be written as $\Gamma_{ij,k\ell} = \frac{1}{2}(\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk})$, using Kronecker's $\delta$ function. We also note that Gaussians are closed under linear operations (see, e.g., [2, eq. 2.115]: $p(B) = \mathcal{N}(\overrightarrow{B}; \overrightarrow{B}_0, V)$ implies $p(\Gamma\overrightarrow{B}) = \mathcal{N}(\Gamma\overrightarrow{B}; \overrightarrow{B}_0, \Gamma V \Gamma^\intercal)$). We complete the proof by observing that

$$(\Gamma(W \otimes W)\Gamma^\intercal)_{ij,k\ell} = \sum_{ab,cd} \tfrac{1}{4}(\delta_{ia}\delta_{jb} + \delta_{ib}\delta_{ja})(\delta_{kc}\delta_{\ell d} + \delta_{kd}\delta_{\ell c}) W_{ac} W_{bd} \tag{A.1}$$

$$= \tfrac{1}{4}(W_{ik}W_{j\ell} + W_{i\ell}W_{jk} + W_{jk}W_{i\ell} + W_{j\ell}W_{ik}) \tag{A.2}$$

$$= \tfrac{1}{2}(W_{ik}W_{j\ell} + W_{i\ell}W_{jk}). \qquad \square \tag{A.3}$$

**A.2. Proof for Theorem 2.3.** To be shown: Given $p(B) = \mathcal{N}(\overrightarrow{B}; \overrightarrow{B}_0, W \otimes\!\!\otimes W)$, the posterior from the likelihood $\delta(Y - BS) = \lim_{\Lambda \to 0} \mathcal{N}(Y; (\boldsymbol{I} \otimes S)B, \Lambda)$, with $Y, S \in \mathbb{R}^{N \times M}$ and $\mathrm{rk}(S) = M$, has mean (with $\Delta = Y - B_0 S$)

$$\begin{aligned} B_M = B_0 &+ \Delta(S^\intercal W S)^{-1} W S^\intercal + W S (S^\intercal W S)^{-1} \Delta^\intercal \\ &- W S (S^\intercal W S)^{-1} (S^\intercal \Delta)(S^\intercal W S)^{-1} S^\intercal W \end{aligned} \tag{A.4}$$

and covariance

$$V_M = (W - WS(S^\intercal W S)^{-1} S^\intercal W) \otimes\!\!\otimes (W - WS(S^\intercal W S)^{-1} S^\intercal W). \tag{A.5}$$

We begin with the posterior mean (A.4). From (2.3), it has the form (with the prior covariance $V = W \otimes\!\!\otimes W$)

$$\boldsymbol{B}_0 + V(\boldsymbol{I} \otimes S)[(\boldsymbol{I} \otimes S^\intercal) V (\boldsymbol{I} \otimes S)]^{-1} \overrightarrow{\Delta}. \tag{A.6}$$

A few straightforward steps establish that the $NM \times NM$ matrix to be inverted is indeed invertible for linearly independent columns of $S$, and has elements

$$[(\boldsymbol{I} \otimes S^\intercal) V (\boldsymbol{I} \otimes S)]_{ia,jb} = \tfrac{1}{2}[W_{ij}(S^\intercal W S)_{ab} + (WS)_{ib}(WS)_{ja}]. \tag{A.7}$$

Also, the elements of $V(\boldsymbol{I} \otimes S)$ are

$$[V(\boldsymbol{I} \otimes S)]_{ij,ka} = \tfrac{1}{2}(W_{ik}S_{ja} + W_{jk}S_{ia}). \tag{A.8}$$

So we are searching the unique matrix $X \in \mathbb{R}^{N \times M}$ satisfying

$$\overrightarrow{\Delta} = [(\boldsymbol{I} \otimes S^\intercal) V (\boldsymbol{I} \otimes S)] \overrightarrow{X} = \tfrac{1}{2}(\overrightarrow{WXS^\intercal WS + WSX^\intercal WS}), \tag{A.9}$$

which then gives the posterior as $\frac{1}{2}(WXS^\intercal WSWSX^\intercal W)$. (Because $X$ is rectangular, (A.9) is a generalization of a Lyapunov equation. Standard solutions for such equations do not apply directly.) Instead of just presenting a solution, the following lines show a

constructive proof. We first rewrite (A.9) ($S^\intercal WS$ is invertible because $W$ is positive definite, and $S$ is assumed to be of rank $M$) as

$$2\Delta = WXS^\intercal WS + WSX^\intercal WS, \tag{A.10}$$

$$2W^{-1}\Delta(S^\intercal WS)^{-1} = X + SX^\intercal WS(S^\intercal WS)^{-1}. \tag{A.11}$$

Let $Q\Sigma U^\intercal = S$ be the singular value decomposition of $S$. That is, $Q \in \mathbb{R}^{N\times N}$ and $U \in \mathbb{R}^{M\times M}$ are orthonormal, $\Sigma \in \mathbb{R}^{N\times M}$, consisting of an upper part containing the diagonal matrix $D \in \mathbb{R}^{M\times M}$ and a lower part in $\mathbb{R}^{(N-M)\times M}$ containing on zeros. We will write $Q = [Q_+, Q_-]$, where $Q_+ \in \mathbb{R}^{N\times M}$ is a basis of the preimage of $S$, and $Q_- \in \mathbb{R}^{(N-M)\times M}$ is a basis of the kernel of $S$. Because $S$ is full rank, $D$ is invertible, and we can equivalently write

$$X = QRD^{-1}U \quad \text{with a (generally dense) matrix} \quad R = \begin{pmatrix} R_+ \\ R_- \end{pmatrix} \tag{A.12}$$

($R_+ \in \mathbb{R}^{M\times M}, R_- \in \mathbb{R}^{(N-M)\times M}$). This allows the rewriting of equation (A.11) as

$$2Q^\intercal W^{-1}\Delta(S^\intercal WS)^{-1} = RD^{-1}U^\intercal + Q^\intercal Q\Sigma U^\intercal(UD^{-1}R^\intercal Q^\intercal WS)(S^\intercal WS)^{-1}, \tag{A.13}$$

$$2\begin{pmatrix} Q_+^\intercal W^{-1}\Delta(S^\intercal WS)^{-1}UD \\ Q_-^\intercal W^{-1}\Delta(S^\intercal WS)^{-1}UD \end{pmatrix} = \begin{pmatrix} R_+ + [R^\intercal Q^\intercal WS](S^\intercal WS)^{-1}UD \\ R_- \end{pmatrix},$$

which identifies $R_-$. Noting that $Q_+Q_+^\intercal = Q_+D^{-1}UU^\intercal DQ_+^\intercal = S^+S^\intercal$, we can write

$$R^\intercal Q^\intercal WS = (R_+^\intercal Q_+^\intercal + R_-^\intercal Q_-^\intercal)WS \tag{A.14}$$

$$= (R_+^\intercal Q_+^\intercal + 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}Q_-Q_-^\intercal)WS \tag{A.15}$$

$$= R_+^\intercal Q_+^\intercal WS + 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}(\boldsymbol{I} - Q_+Q_+^\intercal)WS \tag{A.16}$$

$$= R_+^\intercal Q_+^\intercal WS + 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal S - 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}S^+(S^\intercal WS). \tag{A.17}$$

Plugging back into (A.13), using $(S^\intercal WS)^{-1}UD = (Q_+^\intercal WS)^{-1}$, we get

$$\begin{aligned} 2Q_+^\intercal W^{-1}\Delta(S^\intercal WS)^{-1}UD &= R_+ + R_+^\intercal Q_+^\intercal WS(Q_+^\intercal WS)^{-1} \\ &\quad + 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal S(S^\intercal WS)^{-1}UD \\ &\quad - 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}S^+(S^\intercal WS)(S^\intercal WS)^{-1}UD \\ &= R_+ + R_+^\intercal + 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal S(S^\intercal WS)^{-1}UD \\ &\quad - 2DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}S^+UD, \end{aligned} \tag{A.18}$$

$$1/2(R_+ + R_+^\intercal) = Q_+^\intercal W^{-1}\Delta(S^\intercal WS)^{-1}UD + DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}Q_+ \tag{A.19}$$

$$- DU^\intercal(S^\intercal WS)^{-1}\Delta^\intercal S(S^\intercal WS)^{-1}UD. \tag{A.20}$$

We see directly that this is a symmetric matrix, because $S^\intercal\Delta = S^\intercal BS - S^\intercal B_0S = \Delta^\intercal S$. Now, noting that $XS^\intercal + SX^\intercal = Q_+(R_+ + R_+^\intercal)Q_+^\intercal + Q_-R_-Q_+^\intercal + Q_+R_-^\intercal Q_-^\intercal$,

we find

$$\begin{aligned}
\text{(A.21)}\qquad \tfrac{1}{2}(XS^\intercal + SX^\intercal) &= (Q_+Q_+^\intercal W^{-1}\Delta(S^\intercal WS)^{-1}S^\intercal) \\
&\quad - S(S^\intercal WS)^{-1}\Delta^\intercal S(S^\intercal WS)^{-1}S^\intercal \\
&\quad + S(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}Q_+Q_+^\intercal \\
&\quad \cdot (\boldsymbol{I} - Q_+Q_+^\intercal)W^{-1}\Delta(S^\intercal WS)^{-1}S^\intercal \\
&\quad + S(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}(\boldsymbol{I} - Q_+Q_+^\intercal) \\
\text{(A.22)}\qquad &= -S(S^\intercal WS)^{-1}\Delta^\intercal S(S^\intercal WS)^{-1}S^\intercal \\
&\quad + W^{-1}\Delta(S^\intercal WS)^{-1}S^\intercal + S(S^\intercal WS)^{-1}\Delta^\intercal W^{-1}.
\end{aligned}$$

From (A.8), the posterior mean can be written as

$$\text{(A.23)}\qquad B_M = B_0 + \tfrac{1}{2}(WXS^\intercal W + WSX^\intercal W),$$

which is clearly equal to (A.4). To establish the form of the posterior covariance, we make use of the structural similarities between the posterior mean and covariance (see (2.3)), and notice that we have just established

$$\begin{aligned}
\text{(A.24)}\qquad &\sum_{ka,nb}(V\mathcal{S})_{ij,ka}(\mathcal{S}^\intercal V\mathcal{S})^{-1}_{ka,nb}\Delta_{nb} \\
&\qquad = [\Delta(S^\intercal WS)^{-1}S^\intercal W + WS(S^\intercal WS)^{-1}\Delta^\intercal]_{ij} \\
&\qquad\quad - [WS(S^\intercal WS)^{-1}\Delta^\intercal S(S^\intercal WS)^{-1}S^\intercal W]_{ij}.
\end{aligned}$$

So we can simply replace $\Delta_{nb}$ with $(\mathcal{S}^\intercal V)_{nb,k\ell} = \tfrac{1}{2}[W_{nk}(S^\intercal W)_{b\ell} + W_{n\ell}(S^\intercal W)_{bk}]$ and find, after a few lines of simple algebra, the form of (A.5) for the posterior covariance. This completes the proof. □

**A.3. Proof for Lemma 3.2.** To be shown: If the Gram matrix $S^\intercal WS$ is diagonal, then the exact posterior mean $B_M$ after $M$ steps, which is

$$\begin{aligned}
\text{(A.25)}\qquad B_M &= B_0 + \Delta(S^\intercal WS)^{-1}S^\intercal W + WS(S^\intercal WS)^{-1}\Delta^\intercal \\
&\quad - WS(S^\intercal WS)^{-1}(S^\intercal\Delta)(S^\intercal WS)^{-1}S^\intercal W,
\end{aligned}$$

is equal to the rank-2 update of $B_{M-1}$ using the Dennis update

$$\begin{aligned}
\text{(A.26)}\qquad B_M &= B_{M-1} + \frac{(y_M - B_{M-1}s_M)c_M^\intercal + c_M(y_M - B_{M-1}s_M)^\intercal}{c_M^\intercal s_M} \\
&\quad - \frac{c_M s_M^\intercal (y_M - B_{M-1}s_M)c_M^\intercal}{(c_M^\intercal s_M)^2} \qquad \text{for } c_M = Ws_M.
\end{aligned}$$

We first harmonize the notation between the two formulations by writing the elements of the diagonal Gram matrix as $(S^\intercal WS)_{ij} = \delta_{ij}c_i^\intercal s_i =: \delta_{ij}a_i$. With this notation, the posterior mean $B_M$, (A.25) can be written as

$$\text{(A.27)}\qquad B_M = B_0 + \sum_{i=1}^{M}\frac{\Delta_i c_i^\intercal + c_i\Delta_i^\intercal}{a_i} + \sum_{i=1}^{M}\sum_{j=1}^{M}\frac{c_i[\Delta^\intercal S]_{ij}c_j^\intercal}{a_i a_j},$$

which can be written recursively as

(A.28)
$$B_M = B_{M-1} + \frac{\Delta_M c_M^\mathsf{T} + c_M \Delta_M^\mathsf{T}}{a_M} - \sum_{i=1}^{M-1} \frac{c_M [\Delta^\mathsf{T} S]_{Mi} c_i^\mathsf{T} + c_i [\Delta^\mathsf{T} S]_{iM} c_M^\mathsf{T}}{a_M a_i} - \frac{c_M [\Delta^\mathsf{T} S]_{MM} c_M^\mathsf{T}}{a_M a_M}$$

(A.29)
$$= B_{M-1} + \left( \Delta_M - \sum_{i=1}^{M-1} \frac{c_i (\Delta^\mathsf{T} S)_{iM}}{a_i} \right) \frac{c_M^\mathsf{T}}{a_M} + \frac{y_M}{a_M} \left( \Delta_M - \sum_{i=1}^{M-1} \frac{c_i (\Delta^\mathsf{T} S)_{iM}}{a_i} \right) - \frac{c_M (\Delta^\mathsf{T} S)_{MM} c_M^\mathsf{T}}{a_M^2}.$$

On the other hand, the expression $y_M - B_{M-1} s_M$ from (A.26) can be written using (A.27) as

(A.30)
$$y_M - B_{M-1} s_M = y_M - B_0 s_M - \sum_{i=1}^{M-1} \frac{\Delta_i c_i^\mathsf{T} s_M + c_i \Delta_i^\mathsf{T} s_M}{a_i} + \sum_{i=1}^{M-1} \sum_{j=1}^{M-1} \frac{c_i (\Delta^\mathsf{T} S)_{ij} c_j^\mathsf{T} s_M}{a_i a_j}.$$

But since, by assumption, $c_i^\mathsf{T} s_M = 0$ for $i \neq M$, this expression simplifies to

(A.31)
$$y_M - B_{M-1} s_M = y_M - B_0 s_M - \sum_{i=1}^{M-1} \frac{c_i \Delta_i^\mathsf{T} s_M}{a_i} = \Delta_M - \sum_{i=1}^{M-1} \frac{c_i \Delta_i^\mathsf{T} s_M}{a_i}.$$

Similarly, the expression $s_M^\mathsf{T} (y_M - B_{M-1} s_M)$ from (A.26) simplifies to

(A.32)
$$s_M^\mathsf{T} (y_M - B_{M-1} s_M) = s_M^\mathsf{T} y_M - s_M^\mathsf{T} B_0 s_M - \sum_{i}^{M-1} \frac{s_M^\mathsf{T} (\Delta_i c_i^\mathsf{T} + c_i \Delta_i^\mathsf{T}) s_M}{a_i} - \sum_{i}^{M-1} \sum_{j}^{M-1} \frac{s_M^\mathsf{T} c_i [\Delta^\mathsf{T} S]_{ij} c_j^\mathsf{T} s_M}{a_i a_j} = s_M^\mathsf{T} \Delta_M.$$

Reinserting these expressions into (A.26), we see that it equals (A.29), which completes the proof. ∎

**A.4. Proof for Lemma 3.3.** The DFP update is the direct update with the choice $W = B$; and the BFGS update is the inverse update with the choice $W = H$. So the Gram matrix, in both cases, is $S^\mathsf{T} B S = Y^\mathsf{T} H Y = S^\mathsf{T} Y$. The $i, j$th element of this symmetric $M \times M$ matrix is $y_i^\mathsf{T} s_j$. The statement to be shown is that this matrix is diagonal if the line search directions are chosen as

(A.33)
$$s_{i+1} = -\alpha_{i+1} H_{i+1} F_i$$

with the residual (the gradient of the equivalent quadratic optimization objective) $F_i = B x_i - b$. We also assume perfect line searches. First, consider the special case where $j = i + 1$ (i.e., subsequent line searches). Because they are in the Dennis class, the estimates for $H$ (irrespective of whether they were constructed by inverting a direct

estimate or using an inverse estimate directly) fulfill the "quasi-Newton equation" $s_i = H_{i+1}y_i = H_{i+1}(F_i - F_{i-1})$. Thus

$$s_{i+1} = -\alpha_{i+1}(s_i + H_{i+1}F_{i-1}). \tag{A.34}$$

The exact line search along $s_i$ ended when $s_i^\intercal F_i = 0$, so

$$\begin{aligned} y_i^\intercal s_{i+1} &= -\alpha_{i+1}(F_i - F_{i-1})^\intercal(s_i + H_{i+1}F_{i-1}) = -\alpha_{i+1}(y_i^\intercal H_{i+1}F_{i-1} - s_i^\intercal F_{i-1}) \\ &= -\alpha_{i+1}(s_i^\intercal F_{i-1} - s_i^\intercal F_{i-1}) = 0 \end{aligned} \tag{A.35}$$

(the last line follows again because, by the quasi-Newton equation, $s_i = H_j y_i$ for all $j > i$). By symmetry of the Gram matrix, (A.35) also implies $y_{i+1}^\intercal s_i = 0$. We complete the proof inductively: Let $j > i+1$ or $i > j+1$, and assume $y_i^\intercal s_{j-a} = y_{j-a}^\intercal s_i = 0 \; \forall a > 0$. Also, $F_{j-1}$ can be written with a telescoping sum as

$$F_{j-1} = (F_{j-1} - F_{j-2} + F_{j-2} - F_{j-3} + \cdots - F_i + F_i) = \sum_{a=i}^{j-1} y_a + F_i. \tag{A.36}$$

Hence

$$y_i^\intercal s_j = -\alpha_j y_i^\intercal(s_{j-1} + H_j F_{j-1}) \qquad \text{[by definition of Newton's direction]} \tag{A.37}$$

$$= -\alpha_j(0 + y_i^\intercal H_j F_{j-1}) \qquad \text{[by induction hypothesis]} \tag{A.38}$$

$$= -\alpha_j s_i^\intercal F_{j-1} \qquad \text{[by quasi-Newton property]} \tag{A.39}$$

$$= -\alpha_j s_i^\intercal \left[ \sum_{a=j-1}^{i} y_a + F_i \right] \qquad \text{[by (A.36)]} \tag{A.40}$$

$$= -\alpha_j s_i^\intercal F_i \qquad \text{[by induction hypothesis]} \tag{A.41}$$

$$= 0 \qquad \text{[because } i\text{th line search is exact]}. \tag{A.42}$$

This completes the proof. ∎

*Remark.* This also implies $F_i^\intercal s_j = 0$ for $i \neq j$: Assume w.l.o.g. that $i > j$. Then use the telescoping sum of (A.36) to get

$$0 = y_i^\intercal s_j = (F_i - F_{i-1})^\intercal s_j = \left( F_i - \sum_{a=j}^{i-1} y_a - F_j \right)^\intercal s_j = F_i^\intercal s_j. \tag{A.43}$$

**Acknowledgments.** The author would like to thank Maren Mahsereci and Martin Kiefel for helpful discussions both prior to and during the preparation of this manuscript. He is particularly grateful to Maren Mahsereci for helpful discussions about details of the symmetric basis in Theorem 2.3. In addition, the author is thankful for comments from the two anonymous reviewers, particularly for pointing out relevant previous work.

## REFERENCES

[1] F. Alizadeh, J-P. A. Haeberley, and M. L. Overton, *Primal-dual interior-point methods for semidefinite programming: Convergence rates, stability and numerical results*, SIAM J. Optim., 8 (1988), pp. 746–768.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York, 2006.
[3] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp., 19 (1965), pp. 577–593.
[4] C. G. Broyden, *Quasi-Newton methods and their application to function minimization*, Math. Comp., 21 (1967), pp. 368–381.
[5] C. G. Broyden, *A new double-rank minimization algorithm*, Notices of the AMS, 16 (1969), p. 670.
[6] L. M. Le Cam, *Convergence of estimates under dimensionality restrictions*, Ann. Statist., 1 (1973), pp. 38–53.
[7] S. D. Conte and C. deBoor, *Elementary Numerical Analysis. An Algorithmic Approach*, 3rd ed., McGraw-Hill, New York, Torornto, 1980.
[8] W. C. Davidon, *Variable Metric Method for Minimization*, Technology report, Argonne National Laboratories, Lemont, IL, 1959.
[9] W. C. Davidon, *Optimally conditioned optimization algorithms without line searches*, Math. Programming, 9 (1975), pp. 1–30.
[10] J. E. Dennis, Jr., *On some methods based on Broyden's secant approximations*, in Numerical Methods for Non-Linear Optimization, Dundee, 1971, Academic Press, London, 1972, pp. 19–34.
[11] J. E. Dennis, Jr. and R. B. Schnabel, *Least change secant updates for quasi-Newton methods*, SIAM Rev., 21 (1979), pp. 443–459.
[12] J. E. Dennis, Jr. and H. F. Walker, *Convergence theorems for least-change secant update methods*, SIAM J. Numer. Anal., 18 (1981), pp. 949–987.
[13] J. E. Dennis, Jr. and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev., 19 (1977), pp. 46–89.
[14] P. Diaconis, *Bayesian numerical analysis*, Statistical Decision Theory And Related Topics, IV, Springer, New York, 1988, pp. 163–175.
[15] P. Diaconis and M. Shahshahani, *The subgroup algorithm for generating uniform random variables*, Probability in Engineering and Informational Sciences, 1 (1987), p. 40.
[16] L. C. W. Dixon, *Quasi-newton algorithms generate identical points*, Math. Programming, 2 (1972), pp. 383–387.
[17] L. C. W. Dixon, *Quasi-Newton techniques generate identical points* II: *The proofs of four new theorems*, Math. Programming, 3 (1972), pp. 345–358.
[18] R. Fletcher, *A new approach to variable metric algorithms*, Comput. J., 13 (1970), p. 317.
[19] R. Fletcher and M. J. D. Powell, *A rapidly convergent descent method for minimization*, Comput. J., 6 (1963), pp. 163–168.
[20] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, Comput. J., 7 (1964), pp. 149–154.
[21] D. M. Gay, *Some convergence properties of Broyden's method*, SIAM J. Numer. Anal., 16 (1979), pp. 623–630.
[22] D. Goldfarb, *A family of variable-metric updates derived by variational means*, Math. Comp., 24 (1970), pp. 23–26.
[23] J. Greenstadt, *Variations on variable-metric methods*, Math. Comp., 24 (1970), pp. 1–22.
[24] P. Hennig, *Fast probabilistic optimization from noisy gradients*, in International Conference on Machine Learning (ICML), 2013.
[25] P. Hennig and M. Kiefel, *Quasi-Newton methods – a new direction*, in International Conference on Machine Learning (ICML), 2012.
[26] P. Hennig and M. Kiefel, *Quasi-Newton methods – a new direction*, J. Mach. Learn. Res., 14 (2013), pp. 834–865.
[27] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Research Nat. Bur. Standards, 49 (1952), pp. 409–436.
[28] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.
[29] D. J. C. MacKay, *Introduction to Gaussian processes*, NATO ASI Series F Computer and Systems Sciences, 168 (1998), pp. 133–166.
[30] H. J. Martinez, *Local and Superlinear Convergence of Structured Secant Methods from the Convex Class*, Tech. Report 88-01, Rice University, Houston, TX, 1988.
[31] R. J. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley and Sons, 2005.
[32] W. Murray, ed., *Numerical Methods for Unconstrained Optimization*, Academic Press, New York, London, 1972.
[33] L. Nazareth, *A relationship between the BFGS and conjugate gradient algorithms and its implications for new algorithms*, SIAM J. Numer. Anal., 16 (1979), pp. 794–800.
[34] J. Nocedal, *Updating quasi-Newton matrices with limited storage*, Math. Comp., 35 (1980),

pp. 773–782.
[35] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer Verlag, New York, 1999.
[36] A. O'Hagan, *Some Bayesian numerical analysis*, Bayesian Statistics, 4 (1992), pp. 345–363.
[37] M. J. D. Powell, *A new algorithm for unconstrained optimization*, in Nonlinear Programming, O. L. Mangasarian and K. Ritter, eds., Academic Press, New York, 1970, pp. 31–65.
[38] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, 2006.
[39] D. F. Shanno, *Conditioning of quasi-Newton methods for function minimization*, Math. Comp., 24 (1970), pp. 647–656.
[40] C. F. Van Loan, *The ubiquitous Kronecker product*, J. Comput. Appl. Math., 123 (2000), pp. 85–100.
[41] F. Zhang, *The Schur Complement and Its Applications*, Vol. 4, Springer, New York, 2005.