Following the approach of [1], assume that a model was fit to a data set, which then changed – some data was removed and some added. Denote added by $A$, removed by $R$, and the intersection of old and new data by $I$. Hence, old data $= I \cup R$, and new data $= I \cup A$. For the sake of simplicity, assume $|A| = |R| = m$; denote size of old (and new) by $n$ (typically $n \gg m$).

A model for the data of any set $D$ is obtained by minimizing

$$\sum_{d \in D} f(\theta, d) + |D|\frac{\lambda}{2}||\theta||^2$$

for a vector of parameters $\theta$. It is assumed here (as opposed to [1]) that $f$ is twice differentiable in $\theta$, and (as in [1]) that it is convex.

Denote the old model $\theta_0$ and the new one $\theta$. The goal is to bound the difference between the two. Denote this difference $\delta$, that is, $\theta = \theta_0 + \delta$.

Evidently $\theta_0$ satisfies

$$\sum_{d \in I} \nabla f(\theta_0, d) + \sum_{d \in R} \nabla f(\theta_0, d) + n\lambda \theta_0 = 0$$

And similarly

$$\sum_{d \in I} \nabla f(\theta_0 + \delta, d) + \sum_{d \in A} \nabla f(\theta_0 + \delta, d) + n\lambda (\theta_0 + \delta) = 0$$

Next, use the identity, valid for any $g$ (note that $x, h$ below are vectors):

$$\nabla g(x + h) - \nabla g(x) = \left( \int_0^1 H_f(x + th) dt \right) h$$

(this is a generalization of the mean-value theorem for multi-variate vector functions, i.e. from $\mathbb{R}^p$ to $\mathbb{R}^q$ for $p, q > 1$. $H$ denotes the Hessian, so the integral is an integral over matrices. Note that the last operation is matrix times vector).

Applying Eq. 4 to Eq. 3 yields (note that $\sum_{d \in R} \nabla f(\theta_0, d)$ is added and subtracted):

$$\sum_{d \in I} \nabla f(\theta_0, d) + \sum_{d \in R} \nabla f(\theta_0, d) + \lambda n \theta_0 +$$

$$\sum_{d \in A} \nabla f(\theta_0, d) - \sum_{d \in R} \nabla f(\theta_0, d) + \left( \int_0^1 \sum_{d \in I \cup A} H_f(\theta_0 + t\delta) dt \right) \delta + \lambda n \delta = 0$$
But, from Eq. 3, the first three terms in Eq. 5 sum to 0, hence

\[
\delta = \left( \lambda_n I + \int_0^1 \sum_{d \in I \cup A} H_f(\theta_0 + t\delta) dt \right)^{-1} \left( \sum_{d \in I} \nabla f(\theta_0, d) - \sum_{d \in A} \nabla f(\theta_0, d) \right)
\]  

(6)

where \( I \) denotes the identity matrix with the same dimension as \( \theta \).

This expression resembles the one in [1], with the addition of the Hessian term. Also, it allows to derive the same result as in [1]: \( \theta \) is contained in the sphere centered at \( \theta_0 + \frac{\Delta}{2} \) and radius \( \frac{||\Delta||}{2} \), where \( \Delta \) is defined as

\[
\frac{1}{\lambda_n} \left( \sum_{d \in R} \nabla f(\theta_0, d) - \sum_{d \in A} \nabla f(\theta_0, d) \right).
\]

However, the Hessian term can be applied to obtain a tighter bound than in [1]. To see this, denote (Eq. 6):

\[
\lambda_n I + \int_0^1 \sum_{d \in I \cup A} H_f(\theta_0 + t\delta) dt \triangleq H, \quad \sum_{d \in R} \nabla f(\theta_0, d) - \sum_{d \in A} \nabla f(\theta_0, d) \triangleq \nu_0
\]

So we have \( \theta = \theta_0 + H^{-1}\nu_0 \). To improve the bound in [1], we use the following lemma:

**Lemma 1** Let \( P \) be a positive definite matrix with smallest eigenvalue equal to \( s \), and let \( u \) be a vector. Then \( P^{-1}u \in B(\frac{u}{2s}, \frac{||u||}{2s}) \) (\( B(z, \rho) \) denotes the solid sphere with center \( z \) and radius \( \rho \)).

**Proof 1**: Let \( P \) be diagonalized by an orthonormal \( U \), so \( P = UD U^T \) for a diagonal \( D \) (recall that all of \( D \)'s elements are positive and \( \geq s \)). We seek to prove that \( ||P^{-1}u - \frac{u}{2s}|| \leq \frac{||u||}{2s} \). Now, \( ||P^{-1}u - \frac{u}{2s}|| = ||UD^{-1}U^T u - \frac{u}{2s}|| = ||D^{-1}U^T u - \frac{u}{2s}||. \) Denote \( U^T u = z; \) clearly \( ||z|| = ||u|| \), so we need to prove that \( ||D^{-1}z - \frac{z}{2s}|| \leq \frac{||z||}{2s} \). It suffices to prove that, for every index \( i \),

\[
|\frac{z_i}{D_i} - \frac{z_i}{2s}| \leq \frac{|z_i|}{2s}, \text{ or } |\frac{1}{D_i} - \frac{1}{2s}| \leq \frac{1}{2s}. 
\]

The proof follows easily by separating to two case: \( D_i \geq 2s \) or \( 2s \geq D_i \geq s \).

Plugging this results into Eq. 6, we obtain the following, which tightens the bound in [1]:

Define \( g = \sum_{d \in R} \nabla f(\theta_0, d) - \sum_{d \in A} \nabla f(\theta_0, d) \). Then \( \theta \) is contained in the sphere with center \( \theta_0 + \frac{g}{2(h + \lambda_n)} \) and radius \( \frac{||g||}{2(h + \lambda_n)} \), where \( h \) is the smallest eigenvalue of \( \int_0^1 \sum_{d \in I \cup A} H_f(\theta_0 + t\delta) dt \). Due to the concavity of the smallest eigenvalue, \( h \) is bounded from below by the smallest eigenvalue of \( \sum_{d \in I \cup A} H_f(\theta_0 + t\delta) \) for every \( 0 \leq t \leq 1 \).

2
The problem, of course, is that $\delta$ appears in both sides of the identity. Possibly, this can be overcome by first bounding $\delta$ without using the Hessian term, and then taking the minimal eigenvalue over the set defined by this larger bound, or by using a global bound on the Hessian’s smallest eigenvalue.

**Questions for further research:**

1. For the centralized case, can the size of the Hessian’s smallest eigenvalue be bounded from below, to allow to quickly compute bounds which improve those in [1]?

2. Can the approach outlined here (with the Hessian term), be extended to the distributed case? I think it can (it seems to be related to [2]).

3. How much do we lose by demanding second differentiability of $f$, and can the method be extended to loss functions which are not twice differentiable? I believe “not too much”, since there must be expressions or bounds which can replace the Hessian in this case (such as subgradients can do). There are also cases in which $f$ is already twice differentiable (e.g. logistic regression).
