Homogeneous Second-Order Descent Method for Unconstrained Smooth Optimization

Yinyu Ye

MS&E and ICME, Stanford University

http://www.stanford.edu/~yyye

Chapters 7-10

The Homogeneous Method I

The 2^{nd} order Taylor expansion can be homogenized by adding an auxiliary dimension, e.g.,

$$\begin{aligned} m^k(\mathbf{x}^k) &= f(\mathbf{x}^k) + \nabla f(\mathbf{x}^k)^T(\mathbf{d}^k) + \tfrac{1}{2}(\mathbf{d}^k)^T H^k(\mathbf{d}^k) \\ &= \frac{1}{2t^2}\begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H^k & \mathbf{g}^k \\ (\mathbf{g}^k)^T & 0 \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix} \end{aligned} \tag{1}$$

where $\mathbf{g}^k = \nabla f(\mathbf{x}^k),\ \mathbf{d}^k = v/t$. More generally, we can define a homogeneous model to approximate $m^k(\cdot)$ using δ^k:

$$\psi^k(v, t; \delta) := \frac{1}{2}\begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H^k & \mathbf{g}^k \\ (\mathbf{g}^k)^T & \delta^k \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix} \tag{2}$$

2^{nd} order descent directions can be constructed from ψ^k

Consider Homogeneous Second-Order Descent Method (HSODM):

- minimizing ψ^k over the unit ball,

$$
\begin{array}{rll}
[v^k; t^k] & = \arg\min & \frac{1}{2}\begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H^k & \mathbf{g}^k \\ (\mathbf{g}^k)^T & \delta^k \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix} \\
& \text{s.t.} & \|[v; t]\| \le 1.
\end{array} \tag{3}
$$

$F^k = [H^k, g^k; (\mathbf{g}^k)^T, \delta^k]$ is the aggregated matrix.

- let $\mathbf{x}^{k+1} = \mathbf{x}^k - (\eta^k) v^k / t^k$ with a proper stepsize (η^k) (Line-search, backtracking, etc.).

Theorem 1 *If F^k is indefinite,* (3) *is equivalent to the eigenvalue problem*

When F^k is indefinite, $[v^k; t^k]$ is on the sphere of unit ball, that is, $\|[v^k; t^k]\| = 1$. Then the problem reduce to solving $\lambda_{\min}(F^k)$. For example, we can set $\delta^k < 0$, then F^k must be indefinite.

Comparing to the Newton Equation

When H^k is large, we usually use Krylov subspace method to solve the Newton equation,

$$H^k \mathbf{d}^k = -\mathbf{g}^k. \tag{4}$$

If H^k is positive definite, the Conjugate Gradient Method is linearly convergent with dependence on the condition number $\kappa_H := \lambda_{\max}/\lambda_{\min}$.

While (3) can be solved by a different Krylov method: Lanczos method, which depends on a different gap-dependent condition number defined by the minimum and second minimum eigenvalues:

$$\frac{\lambda_{\max}}{\lambda_2 - \lambda_{\min}} \tag{5}$$

when H^k is degenerate ($\lambda_{\min} = 0$), (3) can be more robust λ_2 is separated from $\lambda_{\min}$.

The Lanczos Method for Symmetric Eigenvalue Problems

Different from the conjugate gradient method, the Lanczos method uses the Krylov subspace $\{\mathbf{g}^0, H^k\mathbf{g}^0, ...\}$ to build the *tridiagonalization:*

$$\begin{aligned} \alpha^k &= (q^k)^T H^k q^k \\ r^k &= \left(H^k - \alpha^k I\right) q^k - \beta^{k-1} q^{k-1} \Rightarrow T^k = \\ q^{k+1} &= r^k/\beta^k, \beta^k = \left\|r^k\right\|_2 \end{aligned} \begin{bmatrix} \alpha^1 & \beta^1 & & \cdots & \\ \beta^1 & \alpha^2 & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & \beta^{k-1} \\ & \cdots & & \beta^{k-1} & \alpha^k \end{bmatrix}$$

Then use T^k to approximate eigenvalues. Generally, the Lanczos method does not require H^k to be positive definite. In theory, the convergence depends on the gap: $\frac{\lambda_{\max}}{\lambda_2 - \lambda_{\min}}$; image $\lambda_{\min} = 0$, we still have a finite "condition number".

The Toy Example

We consider n-dimensional Hilbert matrix:

$$H_{ij} = \frac{1}{i+j-1}, i \leq n, j \leq n. \tag{6}$$

Compare the homogeneous model (3) and Newton equation (4) with a perturbation λ using different λ.

$$\tilde{H} = H + \lambda I \tag{7}$$

Larger λ produces better condition number: $\kappa_{\tilde{H}} = (\lambda_{\max} + \lambda)/(\lambda_{\min} + \lambda)$

Let us compare the Krylov subspace methods.

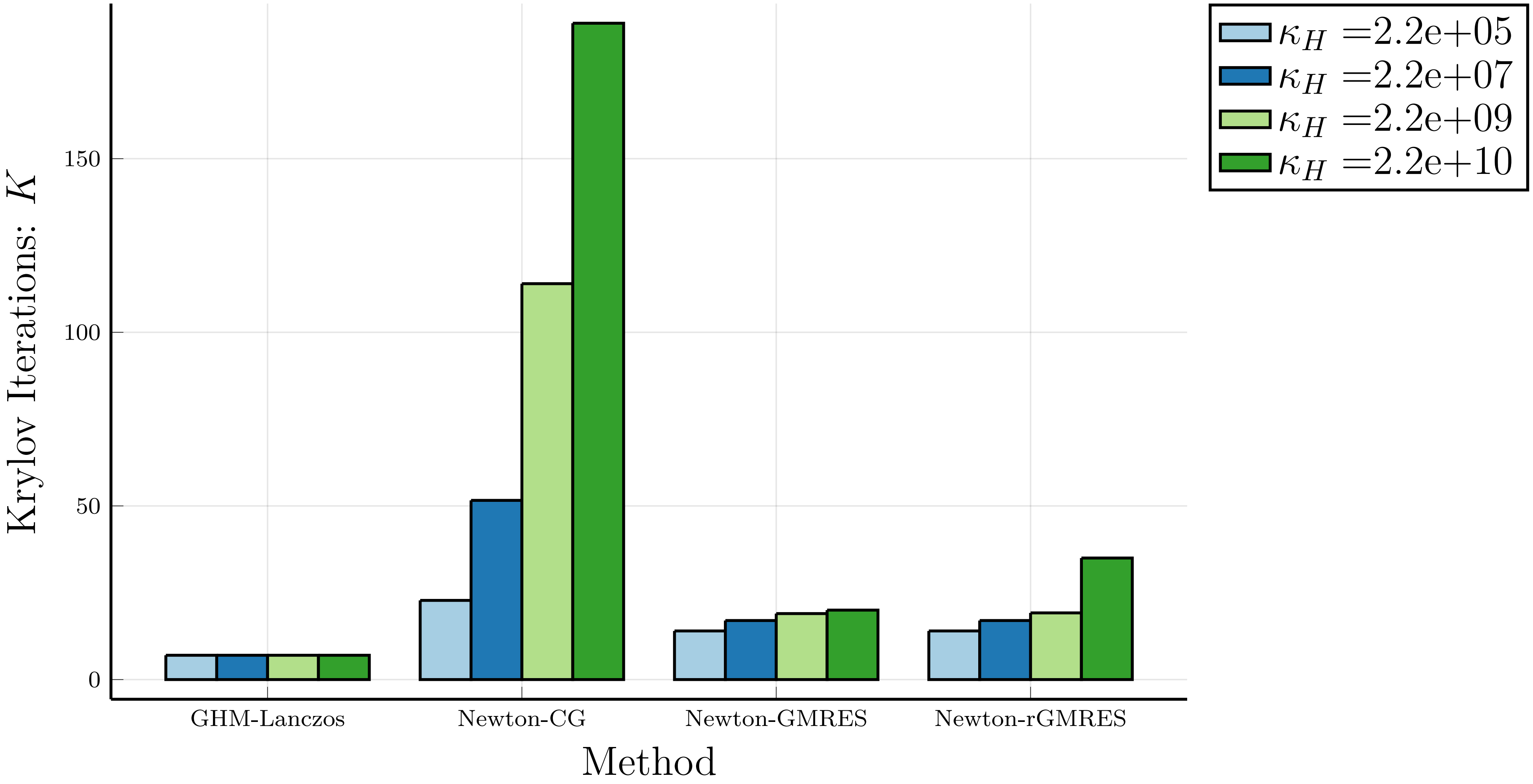


Figure 1: Calculating a Newton-type direction for a perturbed Hilbert matrix

The Lanczos method (GHM-Lanczos) (3) is almost “immune” to large condition numbers. It is also scale-invariant: perturbation does not affect its performance.

Theoretical Guarantee of Solving a Subproblem

Theorem 2 (Kuczyski 92, Golub 13) *The complexity of finding an ϵ-approximate smallest eigenvalue of a symmetric matrix A is (with a high probability)*

- $O(n^2 \cdot \sqrt{\frac{\lambda_{\max}}{\lambda_2 - \lambda_{\min}}} \log(1/\epsilon))$.
- *or* $O(n^2 \cdot \sqrt{\frac{\lambda_{\max}}{\epsilon}} \log(1/\epsilon))$

The *gap-dependent* interpretation (the first one) is particularly meaningful when the matrix is ill-conditioned.

We can see solving the homogeneous model is sometimes easier than solving a Newton equation. We can use this property to construct a SOM based on the homogeneous function ψ^k.

Preliminary Analysis

For illustration, consider a vanilla HSODM. We set $\delta^k \equiv -\sqrt{\epsilon}$ in the homogeneous model (3), consider the eigenvalue problem:

$$[v^k; t^k] = \arg \min_{\|[v;t]\| \le 1} [v; t]^T \begin{bmatrix} H^k & \mathbf{g}^k \\ (\mathbf{g}^k)^T & -\sqrt{\epsilon} \end{bmatrix} [v; t]. \tag{8}$$

Take $\mathbf{d}^k = v^k / t^k$ (if $t^k = 0$ then simply $\mathbf{d}^k = -v^k$). Restrict the step to some $\|(\eta^k)\mathbf{d}^k\| = \Delta^k$:

$$\mathbf{x}^{k+1} = \mathbf{x}^k - (\eta^k)\mathbf{d}^k. \tag{9}$$

A Vanilla HSODM: Overview

The first-order condition for (8):

$$\begin{aligned} \begin{bmatrix} H^k + \theta^k \cdot I & \mathbf{g}^k \\ (\mathbf{g}^k)^T & -\delta^k + \theta^k \end{bmatrix} \begin{bmatrix} v^k \\ t^k \end{bmatrix} &= 0, \\ \|[v^k; t^k]\| &= 1. \end{aligned} \tag{10}$$

and the second-order condition

$$F^k + \theta^k I = \begin{bmatrix} H^k + \theta^k \cdot I & \mathbf{g}^k \\ (\mathbf{g}^k)^T & -\delta^k + \theta^k \end{bmatrix} \succeq 0 \tag{11}$$

Note that the above equations are conditions for global optimal solutions.

- Since it is a ball-constrained QP, we know the global optimal solution satisfy (10) and (11) except that one have the complementarity: $\theta^k \cdot (\|[v^k; t^k]\| - 1) = 0.$

  This means we must justify $[v^k; t^k]$ is not an "interior" point.

- Since the second diagonal term $\delta^k = -\sqrt{\epsilon} < 0$, then F_k must be indefinite, and $\theta^k > 0$. This implies $\|[v^k; t^k]\| - 1) = 0$ holds.

- In this case, the homogeneous model can be solved as an eigenvalue problem:

$$\lambda_{\min}(F^k) := \min_{\|[v;t]\|=1} [v; t]^T \begin{bmatrix} H^k & \mathbf{g}^k \\ (\mathbf{g}^k)^T & -\sqrt{\epsilon} \end{bmatrix} \tag{12}$$

Preliminary Analysis

We now embark on the convergence analysis of a preliminary HSODM for functions with M-Lipschitz second derivatives. Recall $f(\mathbf{x})$ is has M-Lipschitz Hessian if

$$\|\nabla^2 f(\mathbf{x}^{k+1}) - \nabla^2 f(\mathbf{x}^k)\| \leq M\|\mathbf{x}^{k+1} - \mathbf{x}^k\| \tag{13}$$

Similar to the spherical constrained "trust-region" method, we have to show the homogeneous model produces sufficient decrease at each $\mathbf{x}^k$.

Theorem 3 *Suppose that $f(\mathbf{x})$ is second-order Lipschitz continuous. If $(\eta^k) \leq 1$, we have*

$$f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) \leq -\frac{\Delta^2}{2}\delta + \frac{M}{6}\Delta^3. \tag{14}$$

Basically, it is possible when the step $\|d^k\| = \|v^k/t^k\|$ is sufficiently "big". If not, we may conclude it is almost a second-order stationary point.

Proof: $(\eta^k) \leq 1$ implies $\|d^k\| = \|v^k/t^k\|$ is sufficiently "big", in this case $t^k = 0$ can happen. If $t^k \neq 0$, then

$$
\begin{aligned}
f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) &= f(\mathbf{x}^k + (\eta^k)\mathbf{d}^k) - f(\mathbf{x}^k) \\
&\leq (\eta^k) \cdot (\mathbf{g}^k)^T \mathbf{d}^k + \frac{(\eta^k)^2}{2} \cdot (\mathbf{d}^k)^T H^k \mathbf{d}^k + \frac{M}{6}(\eta^k)^3 \|\mathbf{d}^k\|^3 \\
&\leq -\theta^k \cdot \frac{(\eta^k)^2}{2} \|\mathbf{d}^k\|^2 + \frac{M}{6}(\eta^k)^3 \|\mathbf{d}^k\|^3 \\
&\leq -\frac{\Delta^2}{2}\sqrt{\epsilon} + \frac{M}{6}\Delta^3,
\end{aligned}
$$

where the dual solution $\theta^k \geq -\delta^k \equiv \sqrt{\epsilon}$ (We know $F^k + \theta^k I \succeq 0$).

Othewise if $t^k = 0$,

$$\begin{aligned} f(\mathbf{x}^{k+1}) - f(\mathbf{x}^k) &= f(\mathbf{x}^k + (\eta^k)\mathbf{d}^k) - f(\mathbf{x}^k) \\ &\leq (\eta^k) \cdot (\mathbf{g}^k)^T \mathbf{d}^k + \frac{(\eta^k)^2}{2} \cdot (\mathbf{d}^k)^T H^k \mathbf{d}^k + \frac{M}{6}(\eta^k)^3 \|\mathbf{d}^k\|^3 \\ &= \Delta \cdot (\mathbf{g}^k)^T v^k + \frac{\Delta^2}{2} \cdot (v^k)^T H^k v^k + \frac{M}{6}\Delta^3 \|v^k\|^3 \\ &= -\theta^k \cdot \frac{\Delta^2}{2} \|v^k\|^2 + \frac{M}{6}\Delta^3 \|v^k\|^3 \\ &\leq -\frac{\Delta^2}{2}\sqrt{\epsilon} + \frac{M}{6}\Delta^3 \qquad \blacksquare \end{aligned}$$

In trust-region type methods, $t^k = 0$ is referred to as so-called "hard case". This happens only when $\mathbf{g}^k$ is perpendicular to the eigenspace $\mathcal{S}_{\min}(H^k)$ of the smallest eigenvalue.

When the step $\|d^k\| = \|v^k/t^k\|$ (t^k is large), intuitively the second diagonal "dominates" F^k, and should be almost positive semidefinite.

Lemma 1 (Zhang et al. 2022) *If $\|d_k\| \leq \Delta \leq \sqrt{2}/2$, then we have*

$$\|g_k\| \leq 2(L+\delta)\Delta. \tag{15}$$

If so, we choose the full-step $\eta_k = 1$ and $\mathbf{x}^{k+1} = \mathbf{x}^k + \mathbf{d}^k$. We conclude it is a second-order stationary point (SOSP).

Theorem 4 (Zhang et al. 2022) *If $g_k \neq 0$, and $\|d_k\| \leq \Delta$, then let $\eta_k = 1$, we have*

$$\|g_{k+1}\| \leq 2(L+\delta)\Delta^3 + \frac{M}{2}\Delta^2 + \delta\Delta, \tag{16}$$

$$H_{k+1} \succeq -\left(2(L+\delta)\Delta^2 + M\Delta + \delta\right) I \tag{17}$$

We leave these results since they are quite technical; refer to Zhang et al. 2022 if interested.

Preliminary Analysis

In summary, we can use a similar strategy in the spherical constrained "trust-region" method. For example, we can set $\Delta = 2\sqrt{\epsilon}/M$,

- if η^k is small (the produced step $\mathbf{d}^k$ is large), the decrease is guaranteed in $\Omega(\epsilon^{1.5})$; otherwise $\eta^k \leq 1$, we conclude the gradient is small.
- This produces an overall iteration complexity,

$$O((f^0 - f_{\inf})\epsilon^{-1.5}) \tag{18}$$

to an ϵ-approximate SOSP: $\|\mathbf{g}^{k+1}\| \leq O(\epsilon), \lambda_{\min}(H^{k+1}) \geq -\Omega(\sqrt{\epsilon})$.

In the vanilla HSODM, we use a predefined Δ and the a priori knowledge of M. A practical version can utilize line-search methods to adaptively find stepsize η^k.

More Choices in a Homogeneous Framework

The "homogenization" technique can be adjusted for more problems. See the following *generalized homogeneous model* (GHM):

$$F^k := \begin{bmatrix} H^k & \phi^k \\ (\phi^k)^T & \delta^k \end{bmatrix} \tag{19}$$

where $\phi^k \in \mathbf{R}^n$ is a vector. We set δ^k adaptively, and $\mathbf{g}^k$ is not an only option for ϕ^k. For example, use the "inexact gradient", $\|\phi^k - \mathbf{g}^k\| \leq \epsilon$. Recall a Path-Following Method $\mu \to 0$,

$$\mathbf{x}(\mu) = \arg\min f(x) + \mu\|x\|^2 \tag{20}$$

then $\mathbf{x}(\mu) \to \mathbf{x}^*$ (homotopy). Assume f is β-concordant Lipschitz:

$$\|\nabla f(x+d) - \nabla f(x) - \nabla^2 f(x)d\| \leq \beta \cdot d^T \nabla^2 f(x)d, \tag{21}$$

We can use GHMs as subproblems.

A Homotopy HSODM

At some $\mu > 0$, use the GHM as follows:

$$[v^k; t^k] = \arg\min_{\|[v;t]\| \le 1} [v;t]^T \begin{bmatrix} H^k & \mathbf{g}^k + \mu \mathbf{x}^k \\ (\mathbf{g}^k + \mu \mathbf{x}^k)^T & -\mu \end{bmatrix} [v;t]. \tag{22}$$

Just like an interior-point method, solve a sequence of problems by $\mu \to 0$.

- [inner loop] At each μ, solve the GHM repetitively (22)m and set $\mathbf{x}^{k+1} = \mathbf{x}^k - \eta^k v^k / t^k$
- [outer loop] Once $\|\nabla f(x_k) + \mu \cdot x_k\| \le O(\mu)$, decrease $\mu_+ = \sigma \cdot \mu, 0 < \sigma < 1$

If we always start at $\mathbf{x}^k$ after decreasing μ, the [inner loop] has quadratic rate of convergence; the [outer loop] decreases linearly. We have an $O(\log(1/\epsilon)$ algorithm without strong convexity!

Numerical Illustration for Homotopy HSODM

Logistic regression with L_2 penalty Consider the following logistic regression function with L_2 penalty,

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} \log \left(1 + e^{-b_i \cdot a_i^T x}\right) + \frac{\gamma}{2} \|x\|^2, \tag{23}$$

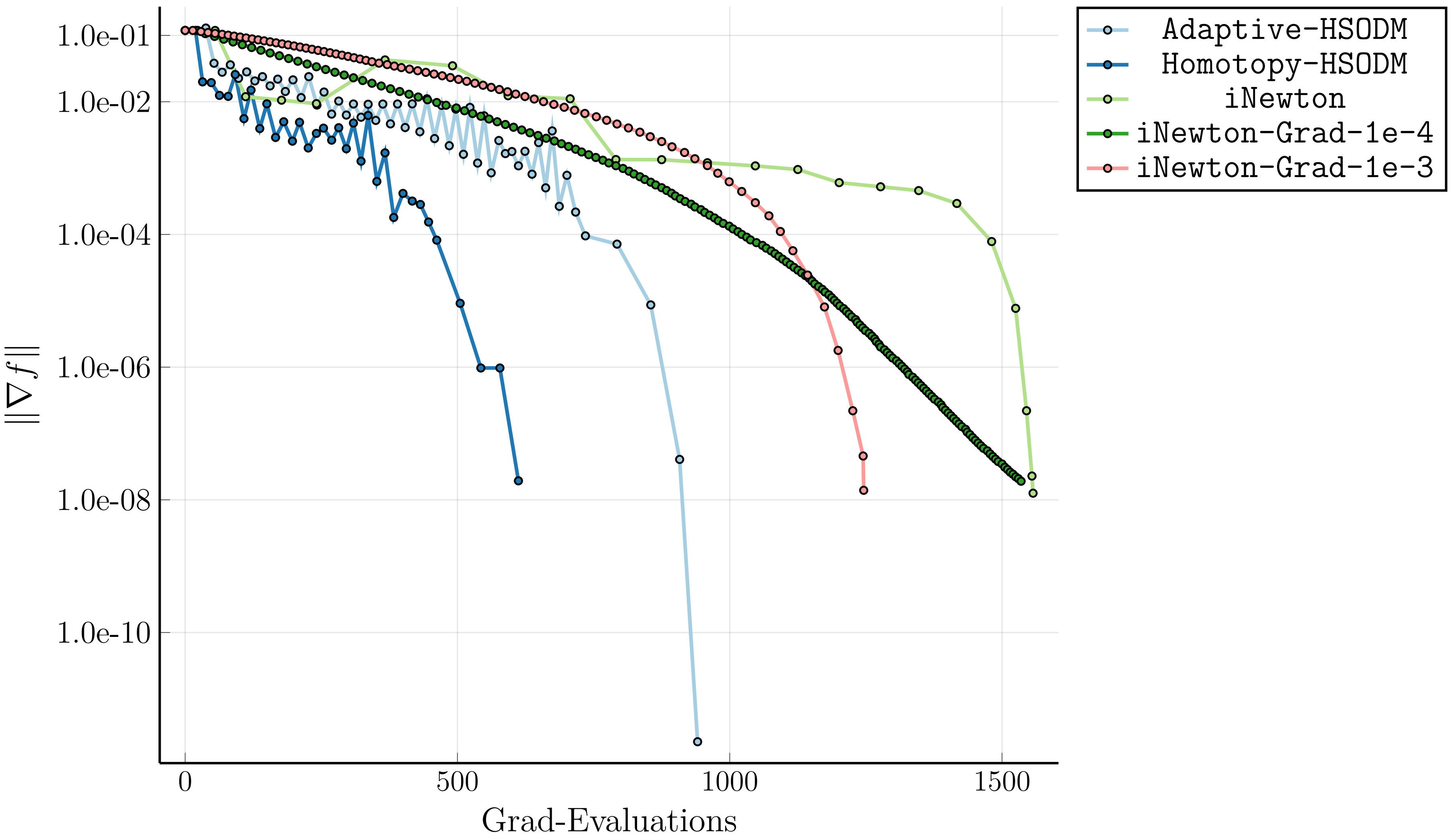
We can show function $f(\mathbf{x})$ is β-concordant Lipschitz. Now we compare SOMs based on Newton equations / Eigenvalue directions, using Krylov subspace methods.

We choose inexact regularized Newton method (`iNewton`) solved by conjugate gradient method:

$$(H^k + \sigma^k \|\mathbf{g}^k\|^{1/2} \cdot I)\mathbf{d}^k = -\mathbf{g}^k \tag{24}$$

(for more details on using "regularization" based on gradient norm; see Mishchenko, SIOPT, 2023).

Logistic Regression **name** := **news20**, n :=1355191, N :=19996



- `Adaptive-HSODM`: adaptive choice for δ^k

- `Homotopy-HSODM` uses 1/3 gradient evaluations/Krylov iterations of a Newton-based SOM.

References

(Original HSODM) Zhang et al, A Homogeneous Second-Order Descent Method for Nonconvex Optimization, arxiv: 2211.08212, (2022)

(Homogeneous Framework) He, Jiang, Zhang et al., Homogeneous Second-Order Descent Framework: A Fast Alternative to Newton-Type Methods, arxiv: 2306.17516, (2023)

(HSODM for Policy Optimization) Tan et al., A Homogenization Approach for Gradient-Dominated Stochastic Optimization, arxiv:2308.10630, (2023)

Software (in Julia): `https://github.com/bzhangcw/DRSOM.jl`