Inexact trust-region algorithms on Riemannian manifolds

Hiroyuki Kasai (The University of Electro-Communications, Japan) and Bamdev Mishra (Microsoft, India)

Problem of interest

- Consider
$$\min_{w \in \mathcal{M}} \left\{ f(w) := \frac{1}{n}\sum_{i=1}^{n} f_i(w) \right\}.$$
  - w is on a Riemannian manifold $\mathcal{M}$ [1].
  - n is number of samples.
- Many promising applications
  - e.g., matrix/tensor completion, subspace tracking.

Contributions

- Propose inexact trust-region algorithms on Riemannian manifolds.
- Propose sub-sampled trust-region algorithms.
- Derive bounds of sample size of sub-sampled gradients and Hessians based on [2, 3, 4].
- Numerical experiments demonstrate significant speed-ups.

Riemannian trust-region (RTR) [1]

- Generalize the Euclidean trust-region (TR).
- Define $\hat{m}_x$ and solve its minima for $\xi \in T_x\mathcal{M}$
$$\hat{m}_x(\xi) = f(x) + \langle \mathrm{grad} f(x), \xi \rangle_x + \frac{1}{2}\langle H(x)[\xi], \xi \rangle_x,$$
  - Approximate m_x of f_x around x, where $m_x = \hat{m}_x \circ R^{-1}$, is obtained from Taylor expansion of pullback of $\hat{f}_x \triangleq f_x \circ R_x$ on tangent space $T_x\mathcal{M}$, where R_x is retraction.
  - $H(x)$ is some symmetric operator on $T_x\mathcal{M}$.
- Find direction and the length of the step, η_k, simultaneously by solving a sub-problem on the vector space $T_x\mathcal{M}$.
- Update iterate x_k
  - $x_k^+ = R_{x_k}(\eta_k)$ is accepted as $x_{k+1} = x_k^+$ when the decrease $\hat{f}_k(x_k) - \hat{f}_k(x_k^+)$ is larger than $\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)$.
  - Otherwise, we accept as $x_{k+1} = x_k$.
- Adjust trust region Δ_k
  - Δ_k is enlarged, unchanged, or shrunk according to the model decrease and the true function decrease.

MATLAB source code

The code compliant to Manopt [5] is available at `https://github.com/hiroyuki-kasai`.

Essential assumptions [2,3,4]

Asm.1. (Manifold and retraction) Consider compact submanifolds in $\mathbb{R}^n$, and second-order retraction.

Asm.2. (Restricted Lipschitz Hessian) There exists $L_H \geq 0$ such that, for all x_k, $\hat{f}_k$ satisfies
$$\left| \hat{f}_k(\eta_k) - f(x_k) - \langle \mathrm{grad} f(x_k), \eta_k \rangle_{x_k} - \frac{1}{2}\langle \eta_k, \nabla^2 \hat{f}_k(0_{x_k})[\eta_k] \rangle_{x_k} \right| \leq \frac{1}{2} L_H \|\eta_k\|^3_{x_k},$$
for all $\eta_k \in T_{x_k}\mathcal{M}$ such that $\|\eta\|_{x_k} \leq \Delta_k$.

Asm.3. (Norm bound on H_k)
$$\|H_k\|_{x_k} \triangleq \sup_{\eta \in T_{x_k}\mathcal{M}, \|\eta\|_{x_k} \leq 1} \langle \eta, H_k[\eta] \rangle_{x_k} \leq K_H.$$

Asm.4. (Approximation error bounds on inexact gradient G_k and Hessian H_k)
$$\|G_k - \mathrm{grad} f(x_k)\|_{x_k} \leq \delta_g,$$
$$\|(H_k - \nabla^2 \hat{f}_k(0_{x_k}))[\eta_k]\|_{x_k} \leq \delta_H \|\eta_k\|_{x_k}.$$

- A typical form in the Euclidean setting, i.e., $\|(H_k - \nabla^2 \hat{f}_k(0_{x_k}))[\eta_k]\|_{x_k} \leq \delta_H \|\eta_k\|^2_{x_k}$ [6], requires that the sample sizes of G_k and H_k need to be *increased* towards convergence.
- Our *relax* form allows the size to be *fixed*.

Asm.5. (Sufficient descent relative to the Cauchy and Eigen directions) [7].

Inexact Hessian and gradient RTR

- Solve approximately a sub-problem $\hat{m}_k(\eta)$ as
$$\begin{cases} f(x_k) + \langle G_k, \eta \rangle_{x_k} + \frac{1}{2}\langle \eta, H_k[\eta] \rangle_{x_k}, & \text{if } \|G_k\|_{x_k} \geq \epsilon_g, \\ f(x_k) + \frac{1}{2}\langle \eta, H_k[\eta] \rangle_{x_k}, & \text{otherwise.} \end{cases}$$
  - Ignoring G_k when $\|G_k\|_{x_k} < \epsilon_g$ is for convergence analysis.

Asm.6. (Gradient and Hessian approx.) Assume $\delta_g < \frac{1-\rho_{TH}}{4}\epsilon_g$ and $\delta_H < \min\{\frac{1-\rho_{TH}}{2}\nu\epsilon_H, 1\}$.

- Need only $\delta_g \in \mathcal{O}(\epsilon_g)$ and $\delta_H \in \mathcal{O}(\epsilon_H)$ [4,Cond.1].

Thm.3.1 (Optimal complexity of **Alg.1**) Consider $0 < \epsilon_g, \epsilon_H < 1$. Suppose **Asms.1**, **2**, and **3** hold. Also, suppose that the inexact Hessian H_k and gradient G_k satisfy **Asm.4** with the approximation tolerance δ_g and δ_H. Suppose that the solution of the sub-problem $\hat{m}_k(\eta)$ satisfies **Asm.5**, and **Asm.6** holds. Then, **Alg.1** returns an (ϵ_g, ϵ_H)-optimal solution in, at most, $T \in \mathcal{O}(\max\{\epsilon_g^{-2}\epsilon_H^{-1}, \epsilon_H^{-3}\})$ iterations.

Inexact RTR algorithm (Alg.1)

Require: $0 < \Delta_{\max} < \infty$, $\epsilon_g, \epsilon_H \in (0,1)$, $\rho_{TH}, \gamma > 1$.

Initialize $0 < \Delta_0 < \Delta_{\max}$, and a starting point $x_0 \in \mathcal{M}$.
for $k = 1, 2, \ldots$ **do**
  Set the approximate (inexact) gradient G_k and H_k.
  if $\|G_k\| \leq \epsilon_g$ and $\lambda_{\min}(H_k) \geq -\epsilon_H$ **then** Return x_k. **end if**
  if $\|G_k\| \leq \epsilon_g$ **then** $G_k = 0$. **end if**
  Calculate $\eta_k \in T_{x_k}\mathcal{M}$ by solving $\eta_k \approx \arg\min_{\|\eta\| \leq \Delta_k} f(x_k) + \langle G_k, \eta \rangle_{x_k} + \frac{1}{2}\langle \eta, H_k[\eta] \rangle_{x_k}$.
  Set $\rho_k = \frac{\hat{f}_k(0_{x_k}) - \hat{f}_k(\eta_k)}{\hat{m}_k(0_{x_k}) - \hat{m}_k(\eta_k)}$.
  if $\rho_k \geq \rho_{TH}$ **then** $x_{k+1} = R_{x_k}(\eta_k)$ and $\Delta_{k+1} = \gamma\Delta_k$.
  else $x_{k+1} = x_k$ and $\Delta_{k+1} = \Delta_k/\gamma$. **end if**
end for
Output x_k.

Sub-sampled RTR (Sub-RTR) for finite-sum problems

- Define the sub-sampled inexact gradient and Hessian for $i \in [n]$ as
$$G_k \triangleq \frac{1}{|\mathcal{S}_g|}\sum_{i \in \mathcal{S}_g} \mathrm{grad} f_i(x_k), \quad H_k \triangleq \frac{1}{|\mathcal{S}_H|}\sum_{i \in \mathcal{S}_H} \mathrm{Hess} f_i(x_k),$$
  - $\mathcal{S}_g, \mathcal{S}_H \subset \{1, \ldots, n\}$ are the set of the sub-sampled indexes, and their sizes are $|\mathcal{S}_g|$ and $|\mathcal{S}_H|$.
- Suppose that $\sup_{x \in \mathcal{M}} \|\mathrm{grad} f_i(x)\|_x \leq K_g^i$ and $\sup_{x \in \mathcal{M}} \|\mathrm{Hess} f_i(x)\|_x \leq K_H^i$ and define $K_g^{\max} \triangleq \max_i K_g^i$ and $K_H^{\max} \triangleq \max_i K_H^i$.

Thm.4.2 (Bounds on sampling size) We define
$$|\mathcal{S}_g| \geq \frac{16(K_g^{\max})^2}{\delta_g^2}\log\frac{2d}{\delta}, \quad |\mathcal{S}_H| \geq \frac{16(K_H^{\max})^2}{\delta_H^2}\log\frac{2d}{\delta}.$$
At any x_k, suppose that sampling is uniform at random to generate $\mathcal{S}_g$ and $\mathcal{S}_H$. Then, we have
$$\Pr(\|G_k - \mathrm{grad} f(x_k)\|_{x_k} \leq \delta_g) \geq 1 - \delta,$$
$$\Pr(\|(H_k - \nabla^2 \hat{f}_k(0_{x_k}))[\eta_k]\|_{x_k} \leq \delta_H \|\eta_k\|_{x_k}) \geq 1 - \delta.$$

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, Optimization Algorithms on Matrix Manifolds. Princeton University Press, 2008.

[2] N. Boumal,P.-A. Absil, C. Cartis. Global rates of convergence for nonconvex optimization on manifolds. IMA J. Numer. Anal., 2018.

[3] P. Xu, F. Roosta-Khorasani, and M. W. Mahoney. Newton-type methods for non-convex optimization under inexact Hessian information. arXiv preprint arXiv:1708.07164, 2017.

[4] Z. Yao, P. Xu, F. Roosta-Khorasani, and M. W. Mahoney. Inexact non-convex Newton-type methods. arXiv preprint arXiv:1802.06925, 2018.

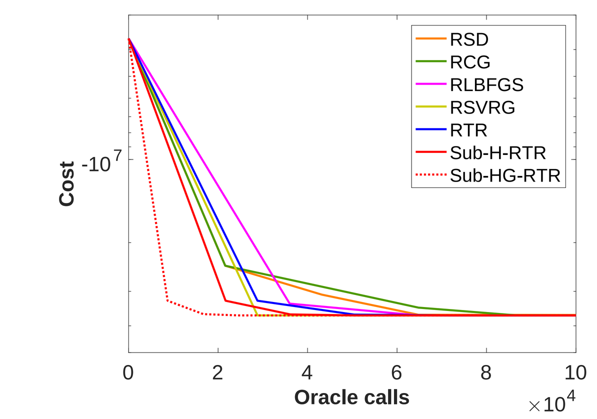
[5] N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. J. Mach. Learn. Res., 15(1):1455-1459, 2014.

[6] C. Cartis, N. I. M. Gould, and P. L. Toint. Adaptive cubic regularisation methods for unconstrained optimization. part I: motivation, convergence and numerical results. Math. Program., 127(2):245-295, 2011.

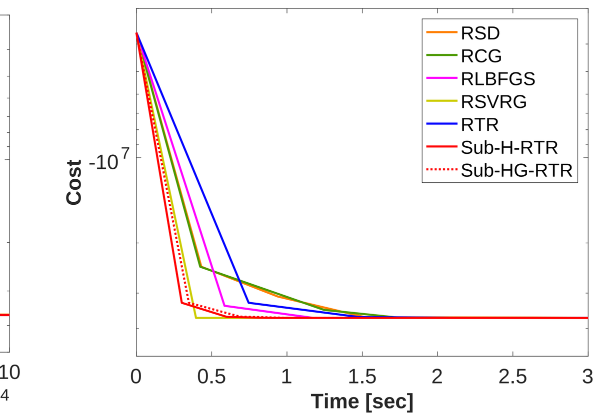
[7] A. R. Conn, N. I. M. Gould, and P. L. Toint. Trust Region Methods. MOS-SIAM Series on Optimization. SIAM, 2000.

Numerical evaluations

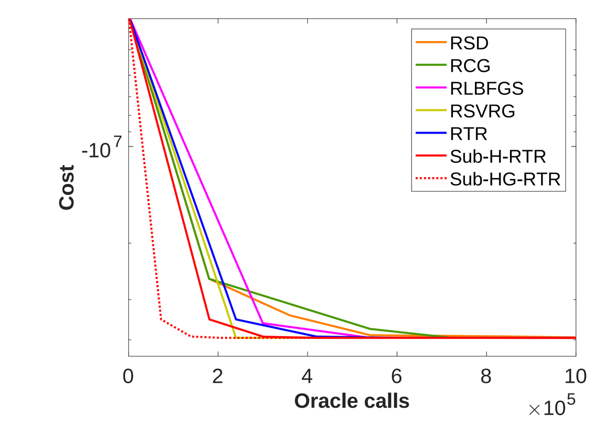
A. ICA problem



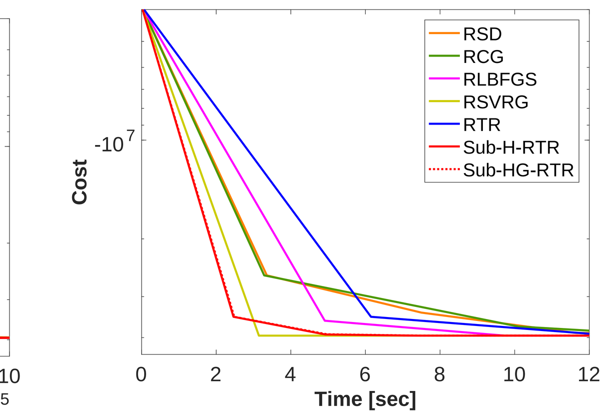
(b-1) Case I2: COIL-100 dataset.



(b-2) Case I2: COIL-100 dataset.

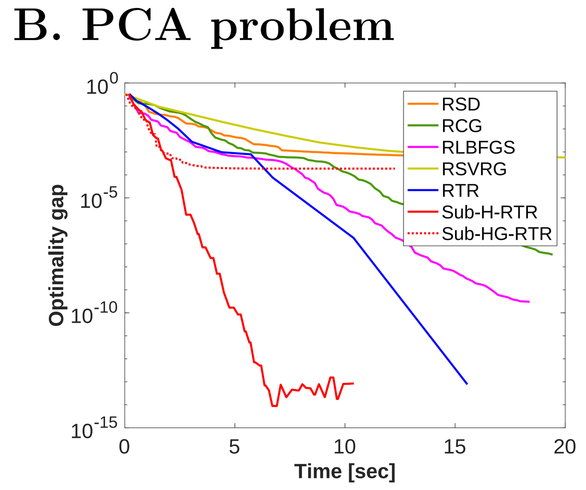


(c-1) Case I3: CIFAR-100 dataset.

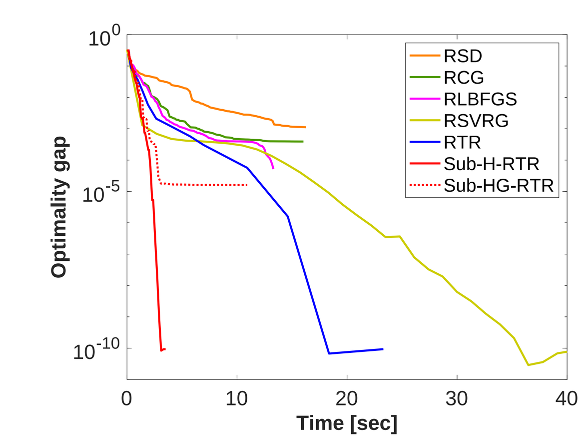


(c-2) Case I3: CIFAR-100 dataset.

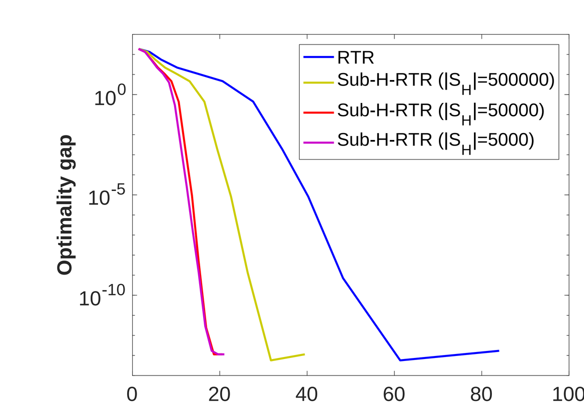
B. PCA problem



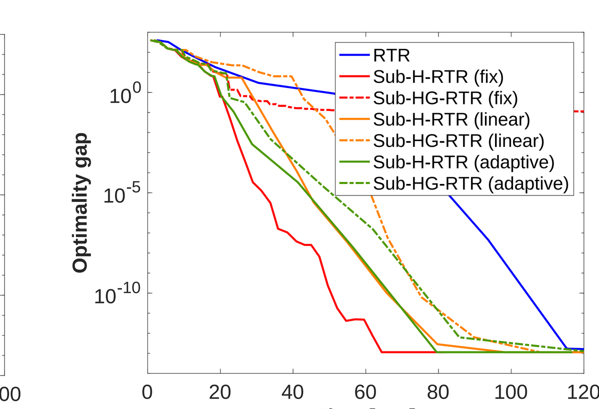
(c) Case P3: MNIST dataset.



(d) Case P4: Covertype dataset.

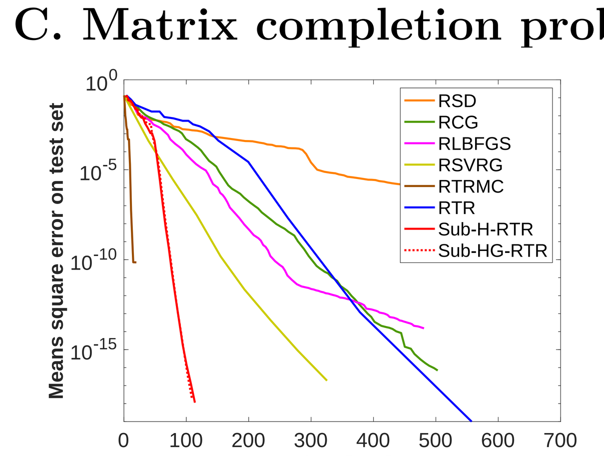


(e) Case P5: Sampling size insensitivity.

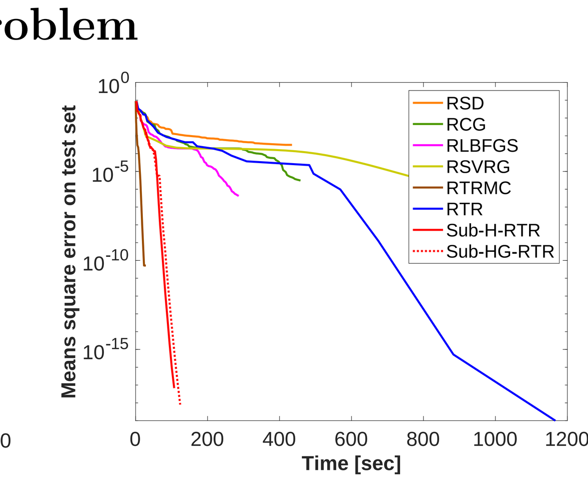


(f) Case P6: Sampling algorithms.

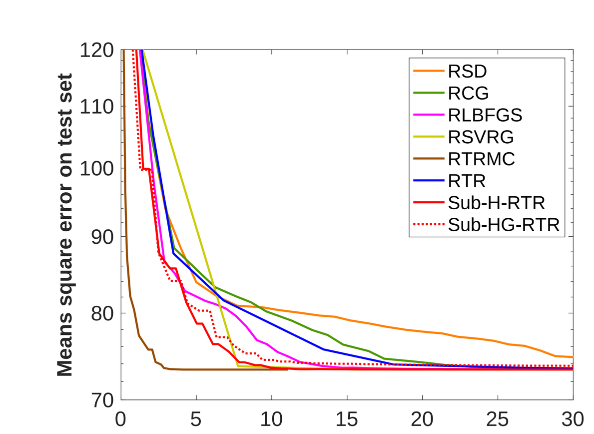
C. Matrix completion problem



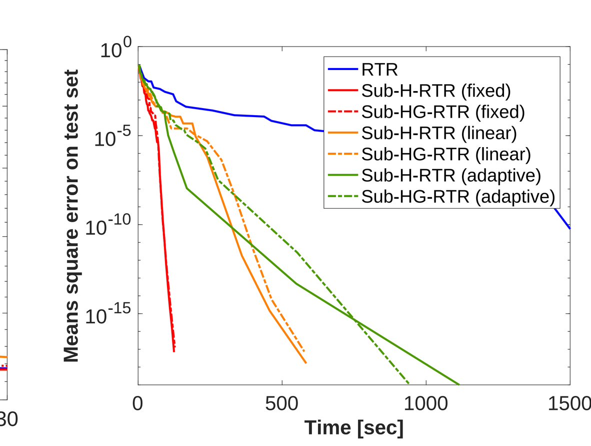
(a-2) Case M1: Synthetic.



(b-2) Case M2: Synthetic (ill-conditioned).



(c-2) Case M3: Jester dataset.



(d) Case M4: Sampling algorithms.