Complexity of two-metric projection method for Bound-constrained Optimization

Hanju Wu
Supervisor: Dr. Yue Xie

Sun Yat-sen University

April 25, 2023

Optimization basics

$$\begin{aligned} \min_{x \in \mathbb{R}^n} \quad & f(x) \\ \text{subject to} \quad & g_i(x) \leq 0, \; i = 1, \ldots, m \\ & h_i(x) = 0, \; i = 1, \ldots, p \end{aligned}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is the objective function.

KKT condition

$$\nabla_x L\left(x^*, \lambda^*\right) = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i\left(x^*\right) + \sum_{i=1}^{p} \lambda_i^* \nabla h_i\left(x^*\right) = 0$$

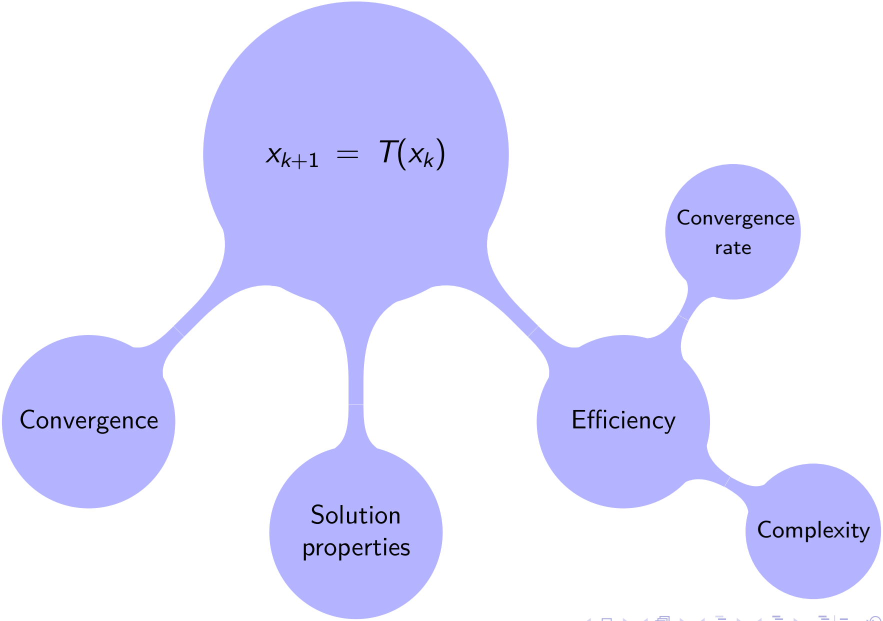
$$g_i\left(x^*\right) \leq 0, \forall i = 1, \ldots, m$$

$$h_i\left(x^*\right) = 0, \forall i = 1, \ldots, p$$

$$\lambda_i^* \geq 0, \forall i = 1, \ldots, m$$

$$\lambda_i^* g_i\left(x^*\right) = 0, \forall i = 1, \ldots, m$$

Optimization basics



What is complexity?

Bound on the amount of computation for an algorithm to solve a class of problems to certain accuracy.

Results for unconstrained nonconvex optimization $\min_{x\in\mathbb{R}^n} f(x)$

Table: Iteration/evaluation complexity for unconstrained optimization

Method	Bound	Assumption	Reference
approx first-order stationary: $\|\nabla f(x)\| \le \epsilon$			
Gradient descent	$\mathcal{O}(\epsilon^{-2})$	∇f L.C.	
approx second-order stationary: $\|\nabla f(x)\| \le \epsilon, \nabla^2 f(x) \succeq -\sqrt{\epsilon}$			
Cubic regularization	$\mathcal{O}(\epsilon^{-3/2})$	$\nabla^2 f$ L.C.	[Nesterov and Polyak, 2006]
Newton-CG	$\mathcal{O}(\epsilon^{-3/2})$	$\nabla^2 f$ L.C.	[Royer et al., 2020]

L.C.: Lipschitz Continuous.

- Iteration denotes outer-loop iteration of the algorithm. It is equivalent to evaluation (gradient, Hessian) complexity but does not necessarily measure the total amount of computational effort.
- We also seek good bounds on the total computation effort.

Bound-Constrained Optimization

$$\min_{x \in \mathbb{R}^n} \; f(x) \quad \text{subject to} \quad x^i \geq 0, i = 1, \ldots, n. \tag{BP}$$

- $f : \mathbb{R}^n \to \mathbb{R}$ is bounded below by f_{low} on the feasible region.

Nonnegative matrix factorization (NMF)

$\min_{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}} \frac{1}{2} \| UV^T - M \|_F^2$ subject to $U \geq 0$, $V \geq 0$.

Assumption

(i). The level set $\mathcal{L}_f(x_0) \triangleq \{x \mid x^i \geq 0, \ \forall i, \ f(x) \leq f(x_0)\}$ *is compact.*
(ii). f *is gradient Lipschitz continuously differentiable on an open convex set* $\mathcal{D}$ *containing* $\mathcal{L}_f(x_0)$ *and all the trial points generated by the algorithm.*

Approximate first-order points

$$\begin{aligned} \nabla_i f(x^*) &= 0 \quad \text{if} \quad x^{*i} > 0, \quad i = 1, \dots, n, \\ \nabla_i f(x^*) &\geq 0 \quad \text{if} \quad x^{*i} = 0, \quad i = 1, \dots, n. \end{aligned} \tag{Exact 1o}$$

Definition (ϵ-1o)

x is an ϵ-1o point if

$$\|S \nabla f(x)\| \leq \epsilon \tag{1}$$

where S is a diagonal matrix s.t. $S[i, i] = x^i$ if $i \in I^+$ and $S[i, i] = 1$, otherwise. Where $I^+ = \{i \mid 0 \leq x^i \leq \bar{\epsilon}, \nabla_i f(x) > 0\}$, $\bar{\epsilon}$ is define as $\bar{\epsilon} = \min\{\epsilon, w\}$, and $w = \|x - \mathcal{P}(x - M \nabla f(x))\|_2$.

- if x is an ϵ-1o point and $\epsilon = 0$, then x is an exact 1o.

Two-Metric Projection [Bertsekas, 1982]

$$x_{k+1} := \mathcal{P}(x_k - \alpha_k D_k \nabla f(x_k)),$$

- $\mathcal{P}(z)$ is the projection onto the feasible region, i.e.

$$[\mathcal{P}(z)]^i = \max\{z^i, 0\},$$

- $D_k \in \mathbb{R}^{n \times n}$, positive definite matrix (but see below!)
- Low per-iteration computational effort.

- For arbitrary D_k, objective function value may not decrease for any $\alpha_k > 0$. (Figure from [Bertsekas, 2014])

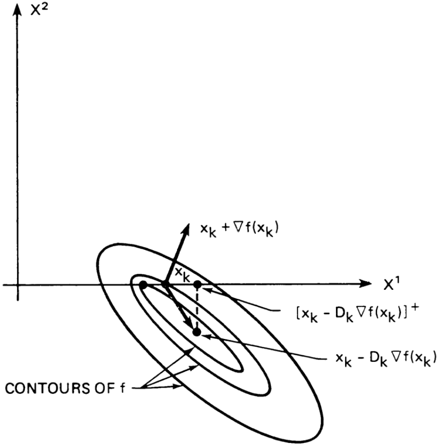


Figure: example

- To ensure descent in f, require

$$D_k = \underbrace{\left(\begin{array}{c|ccc} \bar{D}_k & 0 & \cdots & 0 \\ \hline 0 & d^{r_k+1} & & 0 \\ \vdots & & \ddots & \\ 0 & 0 & & d^n \end{array}\right)}_{I_k^+}$$

$I^+(x_k) \triangleq \{i \mid x_k^i = 0, \nabla_i f(x_k) > 0\}.$

- Unfortunately, the set $I^+(x_k)$ exhibits an undesirable discontinuity at the boundary of the constraint.
- This method is a feasible direction algorithms and has zigzagging or jamming phenomenon.
- For this reason we enlarge the sets $I^+(x_k)$.

Definition

$$I_k^+ = \{ i \mid 0 \le x_k^i \le \epsilon_k, \nabla_i f(x_k) > 0 \}$$

where ϵ_k is define as

$$\epsilon_k = \min\{\epsilon, w_k\}, \quad w_k = \|x_k - \mathcal{P}(x_k - M\nabla f(x_k))\|_2$$

- Under proper assumption, every limit point of a sequence $\{x_k\}$ generate by algorithm 1 is an exact 1o.
- Fast convergence rate (Q-superlinear) when f is convex. Complexity is unknown in nonconvex regime.

for $k = 0, 1, 2, \ldots$ **do**

$w_k = \|x_k - \mathcal{P}_C(x_k - M\nabla f(x_k))\|, \epsilon_k = \min\{\epsilon, w_k\}$

$I_k^+ = \left\{i \mid 0 \le x_k^i \le \epsilon_k, \partial f(x_k)/\partial x^i > 0\right\}$

$p_k = D_k \nabla f(x_k), x_k(\alpha) = \mathcal{P}(x_k - \alpha p_k)$

let m_k is the smallest non negative integer such that

$$f(x_k) - f[x_k(\beta^m)] \ge \sigma \left\{ \beta^m \sum_{i \notin I_k^+} \frac{\partial f(x_k)}{\partial x^i} p_k^i + \sum_{i \in I_k^+} \frac{\partial f(x_k)}{\partial x^i} \left[x_k^i - x_k^i(\beta^m)\right] \right\}$$

stepsize $\alpha_k = \beta^{m_k}$ set $x_{k+1} = x_k(\alpha_k)$

end

Algorithm 1: two-metric projection

Assumption

local optimal point of (BP) satisfies $\forall i \in \mathcal{A}(x^*), \nabla_i f(x^*) > 0$.

Theorem (Convergence rate)

Suppose that f is convex and twice continuously differentiable. Assume that (BP) has a unique optimal solution x^* *satifying assumption, and there exist* $\lambda_{\max} \geq \lambda_{\min} > 0$ *such that*

$$\lambda_{\min}\|z\|^2 \leq z^T \nabla^2 f(x) z \leq \lambda_{\max}\|z\|^2,$$

for all $z \in \mathbb{R}^n$ *and* $x \in \mathcal{L}_f(x_0)$. *Assuming* $D_k = H_k^{-1}$, *where*

$$H_k^{ij} = \begin{cases} 0 & i \neq j \text{ and } i \in I_k^+ \text{ or } j \in I_k^+, \\ \partial^2 f(x_k) / \partial x^i \partial x^j & \text{otherwise.} \end{cases}$$

Then the sequence $\{x_k\}$ *generate by algorithm 1 converges to* x^*, *and the rate of convergence of* $\|x_k - x^*\|$ *is Q-superlinear.*

Approximate 1o Complexity

Theorem (Complexity)

Suppose that there exist $\lambda_{\max} \geq \lambda_{\min} > 0$ such that

$$\lambda_{\min}\|z\|^2 \leq z^T D_k z \leq \lambda_{\max}\|z\|^2,$$

for all $z \in \mathbb{R}^n$ and $k \geq 0$, and D_k is diagonal w.r.t. I_k^+. Then for any $0 < \epsilon < 1$, the two-metric projection method outputs an ϵ-1o in

$$\mathcal{O}(\epsilon^{-3})$$

number of iterations.

Numerical experiments

Nonnegative matrix factorization (NMF)

$$\min_{U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}} \frac{1}{2}\|UV^T - M\|_F^2 \quad \text{subject to } U \geq 0, V \geq 0.$$

Data generation: $M = \bar{U}\bar{V}^T + E$, $\bar{U}$, $\bar{V}$ elements are random and sparse, E is Gaussian noise.

Comparison with specialized solvers

- PNCG: Projected Newton-CG [Xie and Wright, 2021].
- alspgrad: Alternating nonnegative least squares using projected gradient method [Lin, 2007].
- alspnm: Alternating nonnegative least squares using two-metric projection method [Gong and Zhang, 2012].

$m = 300, n = 200$

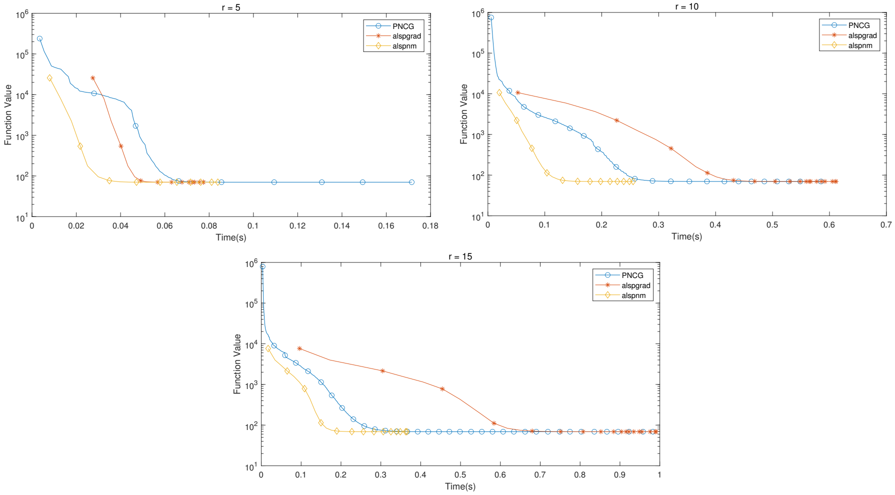


Figure: compare1

large scale

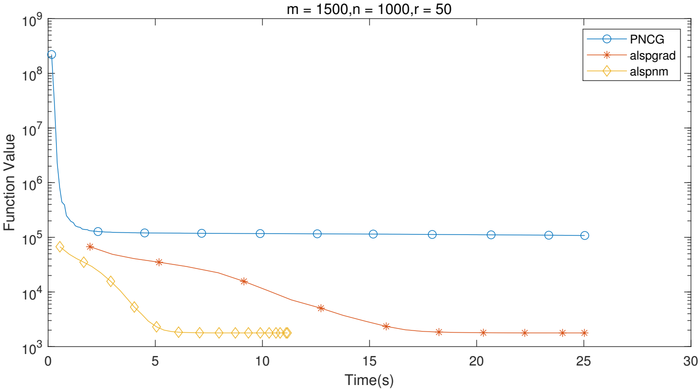


Figure: compare2

- PNCG experiences zigzagging or jamming phenomenon like projected gradient.
- Two-metric projection lowers the objective value quickly.

Summary & generalization

- We present a two-metric projection method featuring both fast convergence rate and good complexity to solve bound-constrained optimization problems.
- Can be generalized to deal with partial two-sided bounds:

$$\min_{x} \; f(x) \quad \text{subject to} \quad 0 \leq x_i \leq u_i, \; \forall i \in \mathcal{I} \subseteq \{1, \ldots, n\}.$$

Future work

- Extension to linear inequality constraints, ℓ_1-norm or ℓ_0-norm constraints.
- Further study on Two-metric projection method and its acceleration.

References I

Bertsekas, D. P. (1982).
Projected newton methods for optimization problems with simple constraints.
SIAM Journal on control and Optimization, 20(2):221–246.

Bertsekas, D. P. (2014).
Constrained optimization and Lagrange multiplier methods.
Academic press.

Gong, P. and Zhang, C. (2012).
Efficient nonnegative matrix factorization via projected newton method.
Pattern Recognition, 45(9):3557–3565.

Lin, C.-J. (2007).
Projected gradient methods for nonnegative matrix factorization.
Neural computation, 19(10):2756–2779.

Nesterov, Y. and Polyak, B. T. (2006).
Cubic regularization of newton method and its global performance.
Mathematical Programming, 108(1):177–205.

References II

Royer, C. W., O'Neill, M., and Wright, S. J. (2020).
A newton-cg algorithm with complexity guarantees for smooth unconstrained optimization.
Mathematical Programming, 180:451–488.

Xie, Y. and Wright, S. J. (2021).
Complexity of a projected newton-cg method for optimization with bounds.
arXiv preprint arXiv:2103.15989.