Pairwise Conditional Gradients without Swap Steps and Sparser Kernel Herding

Kazuma Tsuji (MUFG bank)
Ken'ichiro Tanaka (The University of Tokyo, PRESTO)
Sebastian Pokutta (AISST, ZIB)

ICML2022

CG algorithm

Conditional Gradients (Levitin and Polyak, 1966) are in an important class of first-order methods for constrained convex minimization, i.e., solving

$$\min_{x \in C} f(x) \quad (f : \text{convex}, C \subset \mathbb{R}^d : \text{convex compact region}).$$

Algorithm

1. $v_i = \mathrm{argmax}_{v \in V_C} \langle -\nabla f(\xi_i), v \rangle \quad (C = \mathrm{conv}(V_C))$
2. $\xi_{i+1} = \xi_i + \alpha_i(v_i - \xi_i) = (1-\alpha_i)\xi_i + \alpha_i v_i \ (0 \leq \alpha_i \leq 1)$

PCG and BCG

Pairwise Conditional Gradients (PCG) (Lacoste-Julien and Jaggi, 2015)

- The update manner is $\xi_{i+1} = \xi_i + \alpha_i(v_i - a_i)$
  $(a_i = \mathrm{argmin}_{v \in S_i} \langle -\nabla f(\xi_i), v \rangle$, $S_i = \{v_j\}_{j=1}^{i})$.

Blended Conditional Gradients (BCG) (Braun et al., 2019)

- Add a new vertex v_{i+1} only when the convex coefficients $\{\omega_i\}_{j=1}^{i}$ of $\xi = \sum_{j=1}^{i} \omega_j v_j$ are sufficiently optimized.
- Output sparse solutions.

Table: Theoretical convergence rate (finite-dimensional cases)

	L-smooth	Strongly convex and polytope
PCG	$O(\frac{1}{T})$	$\exp(-c_{\mathrm{P}}T)$
BCG	$O(\frac{1}{T})$	$\exp(-c_{\mathrm{B}}T)$

However, both algorithms suffer in high-dimensional cases. In particular, we cannot guarantee convergence in infinite-dimensional cases !

BPCG algorithm (proposed algorithm)

We propose the following BPCG algorithm. The framework uses that of BCG and the difference is the *local Pairwise step*.

Algorithm Blended Pairwise Conditional Gradients

---

for $t = 0$ to $T - 1$ **do**
  $a_t \leftarrow \mathrm{argmin}_{v \in S_t} \langle -\nabla f(\xi_t), v \rangle$
  $s_t \leftarrow \mathrm{argmax}_{v \in S_t} \langle -\nabla f(\xi_t), v \rangle$
  $v_t \leftarrow \mathrm{argmax}_{v \in V_C} \langle -\nabla f(\xi_t), v \rangle$
  if $\langle \nabla f(\xi_t), a_t - s_t \rangle \geq \langle \nabla f(\xi_t), \xi_t - v_t \rangle$ **then**
    $\xi_{t+1} = \xi_t + \alpha_t (s_t - a_t)$ {local Pairwise step}
  else
    $\xi_{t+1} = \xi_t + \alpha_t (v_t - \xi_t)$ {FW step}
  end if
end for

Using local Pairwise steps, we overcome swap steps (swap of a_t and v_t) which are the bottleneck of PCG in high-dimensional cases.

Theoretical analysis: general smooth case

Theorem

P : convex feasible domain with diameter D (dim P can be ∞)
f : convex and L-smooth.
Let $\{x_i\}_{i=0}^{T} \subset P$ be the sequence given by the BPCG algorithm. Then, it holds that

$$f(x_T) - f(x^*) \leq \frac{4LD^2}{T}.$$

Since the constant factor $4LD^2$ does not depend on the dimension of the domain, we can apply this result to infinite-dimensional cases!

Theoretical analysis: polytopes and strongly convex case

Theorem

P : finite-dimensional polytope with pyramidal width δ and diameter D
f : μ-strongly convex and L-smooth
Consider the sequence $\{x_i\}_{i=0}^{T} \subset P$ obtained by the BPCG algorithm.
Then, it holds that

$$f(x_T) - f(x^*) \leq (f(x_0) - f(x^*)) \exp\left(-c_{f,P}\, T\right),$$

where $c_{f,P} := \frac{1}{2} \min\{\frac{1}{2}, \frac{\mu\delta^2}{4LD^2}\}$.

Compare BPCG to other variants

- BPCG ensures $O(\frac{1}{T})$ convergence in infinite-dimensional cases.
- BPCG ensures linear convergence for strongly convex and polytope cases.
- Moreover, BPCG outputs highly sparse solutions since BPCG inherits the framewrok of BCG.

Table: Theoretical convergence rate

	L-smooth infinite-dimensional domain	Strongly convex, finite-dimensional polytope
CG	$O(\frac{1}{T})$	$O(\frac{1}{T})$
PCG	✗	$\exp(-c_{\mathrm{P}}T)$
BCG	✗	$\exp(-c_{\mathrm{B}}T)$
BPCG	$O(\frac{1}{T})$	$\exp(-c_{\mathrm{BP}}T)$

Numerical experiments (Kernel Herding)

$\mathcal{P}(\Omega)$: all probability measures on $\Omega \in \mathbb{R}^d$

$\mathrm{MMD}(\cdot, \cdot)$: distance between probability measures measured in a Reproducing Kernel Hilbert Space (RKHS) on Ω

Kernel Herding solves the following minimization problem over infinite-dimensional domain $\mathcal{P}(\Omega)$ using a CG manner:

$$\underset{\xi \in \mathcal{P}(\Omega)}{\operatorname{argmin}} \, \mathrm{MMD}^2(\mu, \xi) \quad (\mu \in \mathcal{P}(\Omega)).$$

The output of Kernel Herding is a discrete measure

$$\xi = \sum_{i=1}^{n} \omega_i \delta_{x_i} \quad (\{\omega_i\}_{i=1}^{n} \subset \mathbb{R}, \{x_i\}_{i=1}^{d} \subset \mathbb{R}^d).$$

Using an efficient CG method, we want to derive ξ that approximates μ with small number of nodes n. That is, we want to derive nice sparse soltions.

BPCG for kernel herding

Domain : $\Omega = [-1, 1]^2$, Kernel : Matérn kernel with $\nu = \frac{3}{2}, \frac{5}{2}$.
Optimal rates of the convergence of MMD is $n^{-\frac{5}{4}}, n^{-\frac{7}{4}}$, respectively.

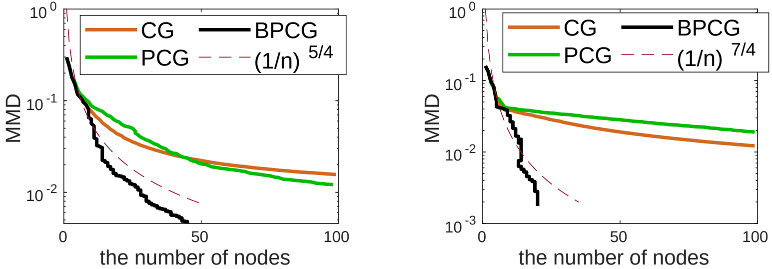

Figure: Matérn kernel ($\nu = 3/2$) (left) and Matérn kernel ($\nu = 5/2$) (right)

Reference I

G. Braun, S. Pokutta, D. Tu, and S. Wright. Blended conditional gradients: the unconditioning of conditional gradients. In *Proceedings of the 36th International Conference on Machine Learning (PMLR)*, volume 97, pages 735–743, 2019.

S. Lacoste-Julien and M. Jaggi. On the global linear convergence of Frank-Wolfe optimization variants. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 496–504. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5925-on-the-global-linear-convergence-of-frank-wolfe-optimi` `pdf`.

E. S. Levitin and B. T. Polyak. Constrained minimization methods. *USSR Computational Mathematics and Mathematical Physics*, 6(5):1–50, 1966.