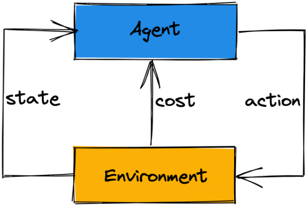
Block Policy Mirror Descent

Guanghui (George) Lan, **Yan Li**, Tuo Zhao

Georgia Institute of Technology

CISS 2022

Markov decision process



Key elements:

- $\mathcal{S}$: state space, finite
- $\mathcal{A}$: action space, finite
- $\mathbb{P}$: transition kernel
- γ: discount factor
- c, h: costs

- **Planning in** $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, c, h)$:

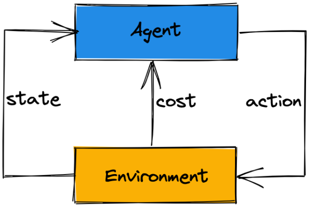
$$\min_{\pi} V^{\pi}(s) := \mathbb{E}_{\pi}[\textstyle\sum_{t=0}^{\infty} \gamma^t \underbrace{(c(s_t, a_t) + h^{\pi}(s_t))}_{\text{policy-dependent cost}} | s_0 = s]$$

- Regularizer $h^{\pi}(s)$ is μ-strongly convex ($\mu \geq 0$) in $\pi(\cdot|s)$ for each s

$$h^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \log \pi(a|s) \qquad \text{(negative entropy)}$$

$$h^{\pi}(s) = 0 \qquad \text{(standard MDP)}$$

Markov decision process



Key elements:

- $\mathcal{S}$: state space, finite
- $\mathcal{A}$: action space, finite
- $\mathbb{P}$: transition kernel
- γ: discount factor
- c, h: costs

- **Planning in $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, c, h)$:**

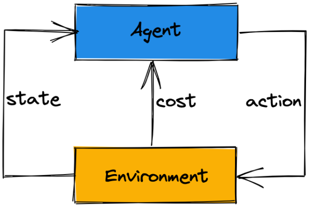
$$\min_{\pi} V^{\pi}(s) := \mathbb{E}_{\pi}[\textstyle\sum_{t=0}^{\infty} \gamma^t \underbrace{(c(s_t, a_t) + h^{\pi}(s_t))}_{\text{policy-dependent cost}} \big| s_0 = s]$$

- Regularizer $h^{\pi}(s)$ is μ-strongly convex ($\mu \geq 0$) in $\pi(\cdot|s)$ for each s

$$h^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \log \pi(a|s) \qquad \text{(negative entropy)}$$

$$h^{\pi}(s) = 0 \qquad \text{(standard MDP)}$$

Markov decision process



Key elements:

- $\mathcal{S}$: state space, finite
- $\mathcal{A}$: action space, finite
- $\mathbb{P}$: transition kernel
- γ: discount factor
- c, h: costs

- **Planning in $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, c, h)$**:

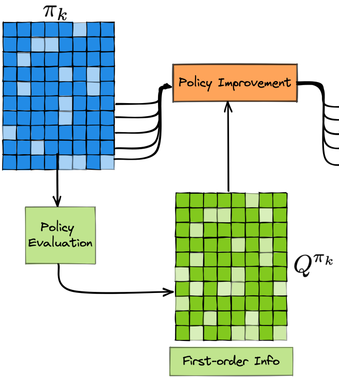
$$\min_{\pi} V^{\pi}(s) := \mathbb{E}_{\pi}[\textstyle\sum_{t=0}^{\infty} \gamma^t \underbrace{(c(s_t, a_t) + h^{\pi}(s_t))}_{\text{policy-dependent cost}} \big| s_0 = s]$$

- Regularizer $h^{\pi}(s)$ is μ-strongly convex ($\mu \geq 0$) in $\pi(\cdot|s)$ for each s

$$h^{\pi}(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \log \pi(a|s) \qquad \text{(negative entropy)}$$

$$h^{\pi}(s) = 0 \qquad \text{(standard MDP)}$$

A Conceptual Recap on Policy Gradient Methods



- Single-objective:

$$f(\pi) = \sum_{s \in \mathcal{S}} \nu^*(s) V^\pi(s)$$

  - ⋆ nonconvex
- Policy evaluation:
  - ⋆ matrix inversion
  - ⋆ TD / simulator
- Policy improvement:
  - ⋆ policy gradient
  - ⋆ natural policy gradient

- Q-function table: $Q^\pi \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ defined as

$$Q^\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^\infty \gamma^t \left(c(s_t, a_t) + h^\pi(s_t) \right) \middle| s_0 = s, a_0 = a \right]$$

Recent developments on Policy Gradient

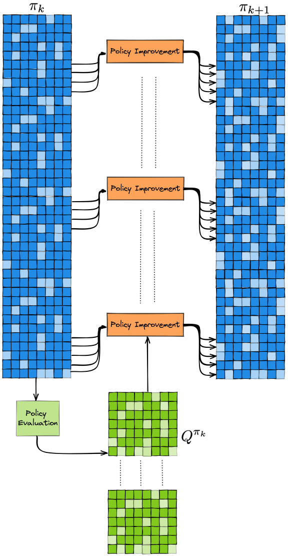
- Possibly even earlier ..
- Even-Dar, Kakade, Mansour '09: $\mathcal{O}(1/\sqrt{T})$ regret of NPG
- Agarwal, Kakade, Lee, Mahajan '19: $\mathcal{O}(1/T)$ of NPG
  - technique inspired by Even-Dar, Kakade, Mansour '09
- Cen, Cheng, Chen, Wei, Chi '20: linear convergence of NPG for entropy regularized MDPs
- Lan '21: (approximate) policy mirror descent
  - linear convergence of NPG/PMD for entropy regularized MDPs
  - linear convergence of APMD for standard MDPs
  - linear convergence of stochastic variants and optimal sample complexity
- Khodadadian, Jhunjhunwala, Varma, Maguluri '21: linear convergence of NPG with adaptive stepsize for standard MDPs

More recently ..

- Li, Lan, Zhao '22: homotopic policy mirror descent
  - linear convergence of standard MDPs, local superlinear convergence
  - last-iterate convergence of the policy
- Xiao '22: linear convergence of NPG/PMD with increasing stepsize

And many more ...

What can be overlooked?



- Both policy improvement and evaluation need to conducted for every state (*batch PG*)
- Per-iteration computation:
  - Evaluation:

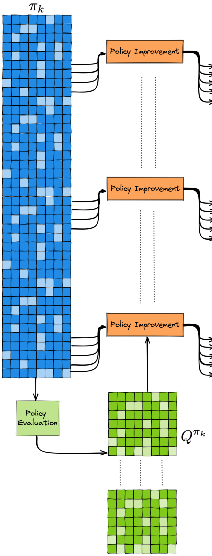
$$\mathcal{O}(\mathrm{MatInv}(|\mathcal{S}|) + |\mathcal{S}|\,|\mathcal{A}|) \quad \text{(known model)}$$
$$\mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|\,/\mathrm{err}) \quad \text{(unknown model)}$$

  - Improvement: $\mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|)$

♠ **Iteration can be very expensive for large state space problem** ♠

Can we design algorithms with cheap iterations, while enjoying similar convergence as batch PG methods?

What can be overlooked?



- Both policy improvement and evaluation need to conducted for every state (*batch PG*)
- Per-iteration computation:
  - Evaluation:

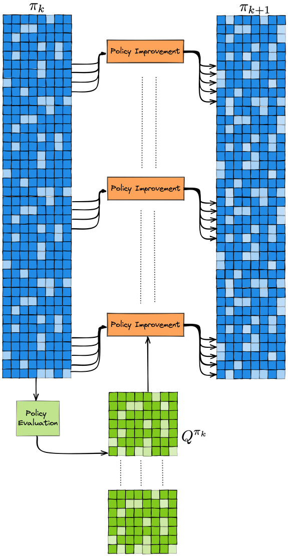
$$\mathcal{O}(\mathrm{MatInv}(|\mathcal{S}|) + |\mathcal{S}|\,|\mathcal{A}|) \quad \text{(known model)}$$
$$\mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|\,/\mathrm{err}) \quad \text{(unknown model)}$$

  - Improvement: $\mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|)$

♠ **Iteration can be very expensive for large state space problem** ♠

Can we design algorithms with cheap iterations, while enjoying similar convergence as batch PG methods?

What can be overlooked?



- Both policy improvement and evaluation need to conducted for every state (*batch PG*)
- Per-iteration computation:
  - Evaluation:

$$\mathcal{O}(\mathrm{MatInv}(|\mathcal{S}|) + |\mathcal{S}|\,|\mathcal{A}|) \quad \text{(known model)}$$
$$\mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|\,/\mathrm{err}) \quad \text{(unknown model)}$$

  - Improvement: $\mathcal{O}(|\mathcal{S}|\,|\mathcal{A}|)$

♠ **Iteration can be very expensive for large state space problem** ♠

Can we design algorithms with cheap iterations, while enjoying similar convergence as batch PG methods?

Presentation Outline

1. **Introduction**
2. **BPMD - Deterministic Variants**
   - Basic BPMD
   - Approximate BPMD
3. **BPMD - Stochastic Variants**
   - Basic SBPMD
   - Approximate SBPMD
4. **Numerical Study**

Part I: Deterministic BPMD Methods

Basic BPMD Method

Idea: blending policy optimization with block coordinate descent

Algorithm The block policy mirror descent (BPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$

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

Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \operatorname*{argmin}_{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s_{i_k},\cdot), p(\cdot|s_{i_k})\rangle + h^p(s_{i_k})\right] + D^p_{\pi_k}(s_{i_k})$$

end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s)\|\pi'(\cdot|s))$
- **Only a single state is updated at each iteration**
  - Evaluating $Q^{\pi_k}(s_{i_k},\cdot)$ reduces to $\mathrm{MatVecMult}(|\mathcal{S}|)$ by exploiting sparse update
  - Cheap policy evaluation and policy improvement
- Can be extended to multi-state update

Basic BPMD Method

Idea: blending policy optimization with block coordinate descent

Algorithm The block policy mirror descent (BPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$

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

Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \operatorname*{argmin}_{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s_{i_k},\cdot), p(\cdot|s_{i_k})\rangle + h^p(s_{i_k})\right] + D^p_{\pi_k}(s_{i_k})$$

end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s)\|\pi'(\cdot|s))$
- **Only a single state is updated at each iteration**
  - Evaluating $Q^{\pi_k}(s_{i_k},\cdot)$ reduces to $\mathrm{MatVecMult}(|\mathcal{S}|)$ by exploiting sparse update
  - Cheap policy evaluation and policy improvement
- Can be extended to multi-state update

Basic BPMD Method

Idea: blending policy optimization with block coordinate descent

Algorithm The block policy mirror descent (BPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$

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

Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \operatorname*{argmin}_{p(\cdot|s_{i_k}) \in \Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s_{i_k}, \cdot), p(\cdot|s_{i_k})\rangle + h^p(s_{i_k})\right] + D^p_{\pi_k}(s_{i_k})$$

end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s) \| \pi'(\cdot|s))$
- **Only a single state is updated at each iteration**
  - Evaluating $Q^{\pi_k}(s_{i_k}, \cdot)$ reduces to $\mathrm{MatVecMult}(|\mathcal{S}|)$ by exploiting sparse update
  - Cheap policy evaluation and policy improvement
- Can be extended to multi-state update

Basic BPMD Method

Idea: blending policy optimization with block coordinate descent

Algorithm The block policy mirror descent (BPMD) method

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$
for $k = 0, 1, \ldots$ **do**
  Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$
  Update policy:
$$\pi_{k+1}(\cdot|s_{i_k}) = \underset{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \eta_k \left[\langle Q^{\pi_k}(s_{i_k},\cdot), p(\cdot|s_{i_k})\rangle + h^p(s_{i_k})\right] + D^p_{\pi_k}(s_{i_k})$$
end for

- $D^{\pi}_{\pi'}(s) := \mathrm{KL}(\pi(\cdot|s)\|\pi'(\cdot|s))$
- **Only a single state is updated at each iteration**
  - Evaluating $Q^{\pi_k}(s_{i_k},\cdot)$ reduces to $\mathrm{MatVecMult}(|\mathcal{S}|)$ by exploiting sparse update
  - Cheap policy evaluation and policy improvement
- Can be extended to multi-state update

Multi-state Variant

Algorithm BPMD with Multi-state Update

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$

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

Sample $\mathcal{B}_k$ uniformly from $[|\mathcal{S}|]$ w.o. replacement

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \operatorname*{argmin}_{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}} \eta_k \left[\langle Q^{\pi_k}(s_{i_k}, \cdot), p(\cdot|s_{i_k})\rangle + h^p(s_{i_k})\right] + D^p_{\pi_k}(s_{i_k}), \ \forall i_k \in \mathcal{B}_k$$

end for

- **Recovers PMD when $\mathcal{B}_k = [|\mathcal{S}|]$**

Convergence of Basic BPMD

Strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu > 0$. Let $\eta_t = \eta$ for all $t \geq 0$, where $\eta > 0$ satisfies $1 + \eta\mu \geq \frac{1}{\gamma}$, then BPMD satisfies

$$\mathbb{E}\left[(f(\pi_k) - f(\pi^*)) + \mu\phi(\pi_k, \pi^*)\right]$$
$$\leq \left(1 - \frac{1-\gamma}{|\mathcal{S}|}\right)^k \left[f(\pi_0) - f(\pi^*) + \frac{\mu}{1-\gamma}\log|\mathcal{A}|\right]$$

- $\mathcal{O}(\frac{|\mathcal{S}|}{1-\gamma}\log(\frac{1}{\epsilon}))$ iterations to find ϵ-optimal policy
- $\mathcal{O}(\frac{|\mathcal{S}|}{(1-\gamma)B}\log(\frac{1}{\epsilon}))$ for multi-state update ($|\mathcal{B}_k| = B$)
  - Recovers the rate of PMD when $\mathcal{B}_k = [|\mathcal{S}|]$
- # policy updates matches best batch PG method (Cen et al. '20, Lan '21)

Convergence of Basic BPMD

Strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu > 0$. Let $\eta_t = \eta$ for all $t \geq 0$, where $\eta > 0$ satisfies $1 + \eta\mu \geq \frac{1}{\gamma}$, then BPMD satisfies

$$\mathbb{E}\left[(f(\pi_k) - f(\pi^*)) + \mu\phi(\pi_k, \pi^*)\right]$$
$$\leq \left(1 - \frac{1-\gamma}{|\mathcal{S}|}\right)^k \left[f(\pi_0) - f(\pi^*) + \frac{\mu}{1-\gamma}\log|\mathcal{A}|\right]$$

- $\mathcal{O}(\frac{|\mathcal{S}|}{1-\gamma}\log(\frac{1}{\epsilon}))$ iterations to find ϵ-optimal policy
- $\mathcal{O}(\frac{|\mathcal{S}|}{(1-\gamma)B}\log(\frac{1}{\epsilon}))$ for multi-state update ($|\mathcal{B}_k| = B$)
  - Recovers the rate of PMD when $\mathcal{B}_k = [|\mathcal{S}|]$
- # policy updates matches best batch PG method (Cen et al. '20, Lan '21)

Convergence of Basic BPMD

Strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu > 0$. Let $\eta_t = \eta$ for all $t \geq 0$, where $\eta > 0$ satisfies $1 + \eta\mu \geq \frac{1}{\gamma}$, then BPMD satisfies

$$\begin{aligned} &\mathbb{E}\left[(f(\pi_k) - f(\pi^*)) + \mu\phi(\pi_k, \pi^*)\right] \\ &\leq \left(1 - \frac{1-\gamma}{|\mathcal{S}|}\right)^k \left[f(\pi_0) - f(\pi^*) + \frac{\mu}{1-\gamma} \log|\mathcal{A}|\right] \end{aligned}$$

- $\mathcal{O}(\frac{|\mathcal{S}|}{1-\gamma}\log(\frac{1}{\epsilon}))$ iterations to find ϵ-optimal policy
- $\mathcal{O}(\frac{|\mathcal{S}|}{(1-\gamma)B}\log(\frac{1}{\epsilon}))$ for multi-state update ($|\mathcal{B}_k| = B$)
  - Recovers the rate of PMD when $\mathcal{B}_k = [|\mathcal{S}|]$
- # policy updates matches best batch PG method (Cen et al. '20, Lan '21)

Convergence of Basic BPMD

Non-strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu = 0$. Let $\eta_t = \eta$ for any $\eta > 0$ and all $k \geq 0$, then BPMD satisfies

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*)\right] \leq \frac{|\mathcal{S}|\left[\eta\big(f(\pi_0) - f(\pi^*)\big) + \log|\mathcal{A}|\right]}{\eta(1-\gamma)k}$$

- $\mathcal{O}(\frac{|\mathcal{S}|}{(1-\gamma)\epsilon})$ number of iteration to find ϵ-optimal policy
- Slow rate! Batch PG can converge linearly

Can we accelerate the sublinear convergence?

Convergence of Basic BPMD

Non-strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu = 0$. Let $\eta_t = \eta$ for any $\eta > 0$ and all $k \geq 0$, then BPMD satisfies

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*)\right] \leq \frac{|\mathcal{S}|\left[\eta\big(f(\pi_0) - f(\pi^*)\big) + \log|\mathcal{A}|\right]}{\eta(1-\gamma)k}$$

- $\mathcal{O}\left(\frac{|\mathcal{S}|}{(1-\gamma)\epsilon}\right)$ number of iteration to find ϵ-optimal policy
- Slow rate! Batch PG can converge linearly

Can we accelerate the sublinear convergence?

The Approximate BPMD Method

Conceptual Idea

Solves a sequence of entropy-regularized MDPs with diminishing regularization

Perturbed MDP - $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, c, h, \tau)$:

- Cost perturbation: $c_\tau^\pi(s,a) = c^\pi(s,a) + \tau D_{\pi_0}^\pi(s)$
- Uniform policy π_0 yields entropy regularization

$$\star \quad Q_\tau^\pi(s,a) = \mathbb{E}_\pi\left[\sum_{t=0}^\infty \gamma^t \left(c^\pi(s_t,a_t) + h^\pi(s_t) + \tau D_{\pi_0}^\pi(s_t)\right) \middle| s_0 = s, a_0 = a\right]$$

The Approximate BPMD Method

Conceptual Idea

Solves a sequence of entropy-regularized MDPs with diminishing regularization

Perturbed MDP - $\mathcal{M}(\mathcal{S}, \mathcal{A}, \mathbb{P}, \gamma, c, h, \tau)$:

- Cost perturbation: $c_\tau^\pi(s,a) = c^\pi(s,a) + \tau D_{\pi_0}^\pi(s)$
- Uniform policy π_0 yields entropy regularization

$$\star \quad Q_\tau^\pi(s,a) = \mathbb{E}_\pi[\textstyle\sum_{t=0}^\infty \gamma^t (c^\pi(s_t,a_t) + h^\pi(s_t) + \tau D_{\pi_0}^\pi(s_t)) | s_0 = s, a_0 = a]$$

The Approximate BPMD Method

Algorithm The Approximate BPMD method

Input: Initial policy π_0, perturbation parameters $\{\tau_k\}_{k\geq 0}$, and stepsizes $\{\eta_k\}_{k\geq 0}$

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

Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \underset{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \eta_k \Big[\big\langle Q^{\pi_k}_{\tau_k}(s_{i_k}, \cdot), p(\cdot|s_{i_k}) \big\rangle + h^p(s_{i_k}) + \tau_k D^p_{\pi_0}(s_{i_k}) \Big] + D^p_{\pi_k}(s_{i_k})$$

end for

- Warm-starting $\mathcal{M}(\mathcal{S}, \mathcal{A}, r, \gamma, \mathbb{P}, \tau_k)$ with $\mathcal{M}(\mathcal{S}, \mathcal{A}, r, \gamma, \mathbb{P}, \tau_{k-1})$

At what rate should τ_k diminish?

The Approximate BPMD Method

Algorithm The Approximate BPMD method

Input: Initial policy π_0, perturbation parameters $\{\tau_k\}_{k\geq 0}$, and stepsizes $\{\eta_k\}_{k\geq 0}$

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

Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \underset{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \eta_k\Big[\big\langle Q^{\pi_k}_{\tau_k}(s_{i_k},\cdot), p(\cdot|s_{i_k})\big\rangle + h^p(s_{i_k}) + \tau_k D^p_{\pi_0}(s_{i_k})\Big] + D^p_{\pi_k}(s_{i_k})$$

end for

- Warm-starting $\mathcal{M}(\mathcal{S}, \mathcal{A}, r, \gamma, \mathbb{P}, \tau_k)$ with $\mathcal{M}(\mathcal{S}, \mathcal{A}, r, \gamma, \mathbb{P}, \tau_{k-1})$

At what rate should τ_k diminish?

Convergence of Approximate BPMD

ABPMD for non-strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose $\mu = 0$ *hold for* h^π. *Let* $l = \lceil \log_{1-(1-\gamma)/|\mathcal{S}|}(1/4) \rceil$, $\tau_t = 2^{-(\lfloor t/l \rfloor + 1)}$, *and* $1 + \eta_t \tau_t = \frac{1}{\gamma}$, *then after* k *iterations,*

$$\begin{aligned}&\mathbb{E}\left[f(\pi_k) - f(\pi^*) + \tau_k \phi(\pi_k, \pi^*)/(1-\gamma)\right] \\ \leq & 2^{-\lfloor \frac{k}{l} \rfloor}\left[f(\pi_0) - f(\pi^*) + 2\log|\mathcal{A}|/(1-\gamma)\right]\end{aligned}$$

- Each regularized MDP solved by $l = \mathcal{O}(|\mathcal{S}| \log_\gamma(\frac{1}{4}))$ iterations
- $\mathcal{O}(\frac{|\mathcal{S}|}{1-\gamma}\log(\frac{1}{\epsilon}))$ iterations to find ϵ-optimal policy
- $\#$ policy updates matches best batch PG method (Lan '21)

Convergence of Approximate BPMD

ABPMD for non-strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose $\mu = 0$ hold for h^π. Let $l = \lceil \log_{1-(1-\gamma)/|\mathcal{S}|}(1/4) \rceil$, $\tau_t = 2^{-(\lfloor t/l \rfloor + 1)}$, and $1 + \eta_t \tau_t = \frac{1}{\gamma}$, then after k iterations,

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*) + \tau_k \phi(\pi_k, \pi^*)/(1-\gamma)\right]$$
$$\leq 2^{-\lfloor \frac{k}{l} \rfloor}\left[f(\pi_0) - f(\pi^*) + 2\log|\mathcal{A}|/(1-\gamma)\right]$$

- Each regularized MDP solved by $l = \mathcal{O}(|\mathcal{S}| \log_\gamma(\frac{1}{4}))$ iterations
- $\mathcal{O}(\frac{|\mathcal{S}|}{1-\gamma} \log(\frac{1}{\epsilon}))$ iterations to find ϵ-optimal policy
- # policy updates matches best batch PG method (Lan '21)

Convergence of Approximate BPMD

ABPMD for non-strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose $\mu = 0$ hold for h^π. Let $l = \lceil \log_{1-(1-\gamma)/|\mathcal{S}|}(1/4) \rceil$, $\tau_t = 2^{-(\lfloor t/l \rfloor + 1)}$, and $1 + \eta_t \tau_t = \frac{1}{\gamma}$, then after k iterations,

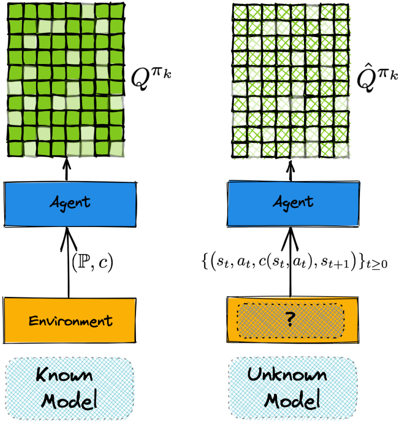
$$\mathbb{E}\left[f(\pi_k) - f(\pi^*) + \tau_k \phi(\pi_k, \pi^*)/(1-\gamma)\right]$$
$$\leq 2^{-\lfloor \frac{k}{l} \rfloor}\left[f(\pi_0) - f(\pi^*) + 2\log|\mathcal{A}|/(1-\gamma)\right]$$

- Each regularized MDP solved by $l = \mathcal{O}(|\mathcal{S}| \log_\gamma(\frac{1}{4}))$ iterations
- $\mathcal{O}(\frac{|\mathcal{S}|}{1-\gamma}\log(\frac{1}{\epsilon}))$ iterations to find ϵ-optimal policy
- # policy updates matches best batch PG method (Lan '21)

Part II: Stochastic BPMD Methods

The Stochastic Variants

Unknown Environment: obtaining exact Q^π can be impractical



Policy update: replace Q^π with sample estimate $Q^{\pi,\xi}$

Basic Stochastic BPMD Method

Algorithm Stochastic BPMD (SBPMD)

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$.

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

Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$.

Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \underset{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \; \eta_k \Big[\Big\langle Q^{\pi_k,\xi_k,i_k}(s_{i_k},\cdot), p(\cdot|s_{i_k}) \Big\rangle + h^p(s_{i_k}) \Big] + D^p_{\pi_k}(s_{i_k})$$

end for

- Construction of Q^{π_k,ξ_k,i_k} can depend on i_k

What conditions should Q^{π_k,ξ_k,i_k} satisfy?

Basic Stochastic BPMD Method

Algorithm Stochastic BPMD (SBPMD)

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$.
for $k = 0, 1, \ldots$ **do**
  Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$.
  Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \operatorname*{argmin}_{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}} \eta_k \Big[\Big\langle Q^{\pi_k,\xi_k,i_k}(s_{i_k},\cdot), p(\cdot|s_{i_k}) \Big\rangle + h^p(s_{i_k}) \Big] + D^p_{\pi_k}(s_{i_k})$$

end for

- Construction of Q^{π_k,ξ_k,i_k} can depend on i_k

What conditions should Q^{π_k,ξ_k,i_k} satisfy?

Conditions on the Noisy Estimate

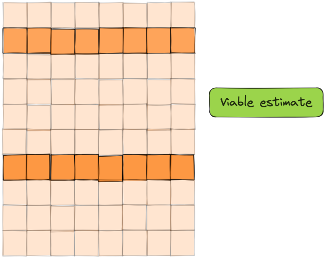
$$\mathbb{E}_{\xi_k|i_k}\left[Q^{\pi_k,\xi_k,i_k}(s_{i_k},\cdot)\right] = \overline{Q}^{\pi_k,i_k}(s_{i_k},\cdot)$$

$$\mathbb{E}_{i_k}\|\overline{Q}^{\pi_k,i_k}(s_{i_k},\cdot) - Q^{\pi_k}(s_{i_k},\cdot)\|_\infty^2 \le v_k^2 \quad \text{(averaged conditional bias)}$$

$$\mathbb{E}_{\xi_k,i_k}\left[\|Q^{\pi_k,\xi_k,i_k}(s_{i_k},\cdot) - Q^{\pi_k}(s_{i_k},\cdot)\|_\infty^2\right] \le \sigma_k^2 \quad \text{(averaged conditional MSE)}$$

Implication for evaluation

$$\hat{Q}^{\pi_k,\xi_k,i_k} - Q^{\pi_k}$$



♣ Okay to have bad estimates for some states - error gets averaged out ♣

Convergence of Basic SBPMD

Strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu > 0$, and $v_t = 2^{-(\lfloor t/l \rfloor + 2)}$, $\sigma_t^2 = 2^{-(\lfloor t/l \rfloor + 2)}$, where $l = \lceil \log_{1-(1-\gamma)/|\mathcal{S}|} \frac{1}{4} \rceil$. Take constant stepsize $\eta_t = \eta > 0$ for all $t \geq 0$, with $1 + \mu\eta = \frac{1}{\gamma}$. Then

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*) + \tfrac{\mu}{1-\gamma}\phi(\pi_k, \pi^*)\right]$$
$$\leq 2^{-\lfloor \frac{k}{l} \rfloor}\left[f(\pi_0) - f(\pi^*) + \tfrac{\mu \log|\mathcal{A}|}{1-\gamma} + \tfrac{5|\mathcal{S}|}{4(1-\gamma)}\left(\tfrac{1}{2\gamma\mu} + 2\right)\right]$$

- Linear convergence with exponentially diminishing noise
  - Typically requires # samples growing exponentially w.r.t. iteration
- $\mathcal{O}(1/\mu k)$ convergence with constant noise
- $\mathcal{O}(1/\sqrt{k})$ convergence when $\mu = 0$

Convergence of Basic SBPMD

Strongly convex regularizers

Theorem (Lan, Li, Zhao '22)

Suppose h satisfies $\mu > 0$, and $v_t = 2^{-(\lfloor t/l \rfloor + 2)}$, $\sigma_t^2 = 2^{-(\lfloor t/l \rfloor + 2)}$, where $l = \lceil \log_{1-(1-\gamma)/|\mathcal{S}|} \frac{1}{4} \rceil$. Take constant stepsize $\eta_t = \eta > 0$ for all $t \geq 0$, with $1 + \mu\eta = \frac{1}{\gamma}$. Then

$$\mathbb{E}\left[f(\pi_k) - f(\pi^*) + \tfrac{\mu}{1-\gamma}\phi(\pi_k, \pi^*)\right]$$
$$\leq 2^{-\lfloor \frac{k}{l} \rfloor}\left[f(\pi_0) - f(\pi^*) + \tfrac{\mu \log|\mathcal{A}|}{1-\gamma} + \tfrac{5|\mathcal{S}|}{4(1-\gamma)}\left(\tfrac{1}{2\gamma\mu} + 2\right)\right]$$

- Linear convergence with exponentially diminishing noise
  - Typically requires # samples growing exponentially w.r.t. iteration
- $\mathcal{O}(1/\mu k)$ convergence with constant noise
- $\mathcal{O}(1/\sqrt{k})$ convergence when $\mu = 0$

Stochastic Approximate BPMD Method

SABPMD for non-strongly convex regularizers

Algorithm Stochastic Approximate BPMD (SABPMD)

Input: Initial policy π_0, and stepsizes $\{\eta_k\}_{k\geq 0}$.
for $k = 0, 1, \ldots$ **do**
  Sample $i_k \sim \mathrm{Unif}([|\mathcal{S}|])$.
  Update policy:

$$\pi_{k+1}(\cdot|s_{i_k}) = \underset{p(\cdot|s_{i_k})\in\Delta_{|\mathcal{A}|}}{\operatorname{argmin}} \eta_k\Big[\Big\langle Q_{\tau_k}^{\pi_k,\xi_k,i_k}(s_{i_k},\cdot), p(\cdot|s_{i_k})\Big\rangle + h^p(s_{i_k}) + \tau_k D_{\pi_0}^p(s_{i_k})\Big] + D_{\pi_k}^p(s_{i_k})$$

end for

- Same as ABPMD, except the stochastic estimate of Q_τ^π

Convergence of Stochastic Approximate BPMD

Theorem (Lan, Li, Zhao '22)

Suppose $\mu = 0$ holds for h. Suppose $\sigma_t^2 = 4^{-(\lfloor t/l \rfloor + 2)}, \upsilon_t = 2^{-(\lfloor t/l \rfloor + 2)}$, where $l = \lceil \log_{1-(1-\gamma)/|\mathcal{S}|}(1/4) \rceil$. Let $\tau_t = 2^{-(\lfloor t/l \rfloor + 1)}$ and $1 + \eta_t \tau_t = \frac{1}{\gamma}$, then

$$\begin{aligned}
&\mathbb{E}\left[f(\pi_k) - f(\pi^*) + \tfrac{\tau_k}{1-\gamma} \phi(\pi_k, \pi^*) \right] \\
\leq & 2^{-\lfloor \frac{k}{l} \rfloor} \left[f(\pi_0) - f(\pi^*) + \tfrac{2 \log |\mathcal{A}|}{1-\gamma} + \tfrac{5|\mathcal{S}|}{4(1-\gamma)} \left(\tfrac{1}{2\gamma} + 1 \right) \right]
\end{aligned}$$

- Linear convergence with exponentially diminishing noise
  - Typically requires # samples growing exponentially w.r.t. iteration

Sample Complexity

Independent Trajectories

Theorem (Lan, Li, Zhao '22)

By using the method of independent trajectories (i.e., assuming generative model) for policy evaluation, the total number of samples of SABPMD can be bounded by

$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log|\mathcal{A}|}{(1-\gamma)^6\epsilon^2}\right)$$

Conditional Temporal Difference

Theorem (Lan, Li, Zhao '22)

By using conditional temporal difference learning (Kotsalis, Lan, Li) for policy evaluation, the total number of samples of SABPMD can be bounded by

$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log^2|\mathcal{A}|\log(1/\epsilon)}{(1-\gamma)^5\epsilon^2}\right)$$

- The dependence on $|\mathcal{S}|$ might be improvable suggested by experiments
  - Can also be improved by using multi-state update
  - Stochastic coordinate descent method has worse sample complexity by a factor of # blocks
- At each iteration, samples required by SABPMD is significantly smaller than batch PG methods

Sample Complexity

Independent Trajectories

Theorem (Lan, Li, Zhao '22)

By using the method of independent trajectories (i.e., assuming generative model) for policy evaluation, the total number of samples of SABPMD can be bounded by

$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log|\mathcal{A}|}{(1-\gamma)^6\epsilon^2}\right)$$

Conditional Temporal Difference

Theorem (Lan, Li, Zhao '22)

By using conditional temporal difference learning (Kotsalis, Lan, Li) for policy evaluation, the total number of samples of SABPMD can be bounded by

$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log^2|\mathcal{A}|\log(1/\epsilon)}{(1-\gamma)^5\epsilon^2}\right)$$

- The dependence on $|\mathcal{S}|$ might be improvable suggested by experiments
  - Can also be improved by using multi-state update
  - Stochastic coordinate descent method has worse sample complexity by a factor of # blocks
- At each iteration, samples required by SABPMD is significantly smaller than batch PG methods

Sample Complexity

Independent Trajectories

Theorem (Lan, Li, Zhao '22)

By using the method of independent trajectories (i.e., assuming generative model) for policy evaluation, the total number of samples of SABPMD can be bounded by

$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log|\mathcal{A}|}{(1-\gamma)^6\epsilon^2}\right)$$

Conditional Temporal Difference

Theorem (Lan, Li, Zhao '22)

By using conditional temporal difference learning (Kotsalis, Lan, Li) for policy evaluation, the total number of samples of SABPMD can be bounded by

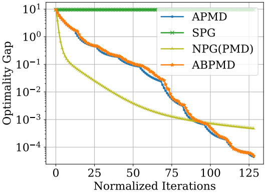
$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|\log^2|\mathcal{A}|\log(1/\epsilon)}{(1-\gamma)^5\epsilon^2}\right)$$

- The dependence on $|\mathcal{S}|$ might be improvable suggested by experiments
  - Can also be improved by using multi-state update
  - Stochastic coordinate descent method has worse sample complexity by a factor of $\#$ blocks
- At each iteration, samples required by SABPMD is significantly smaller than batch PG methods

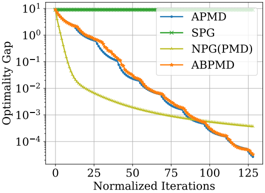
Part III: Numerical Study

Deterministic Setting

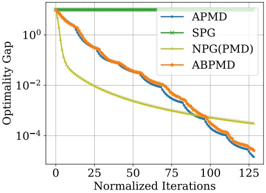
Randomly Generated GridWorld Environments



(a) $|\mathcal{S}| = 225.$



(b) $|\mathcal{S}| = 400.$



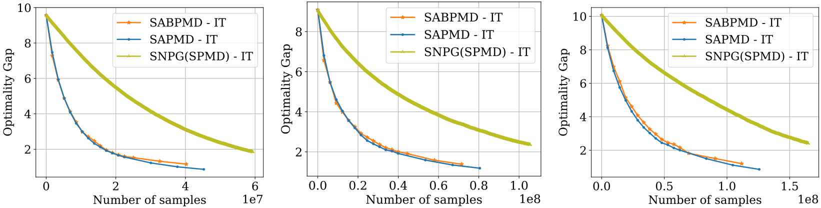
(c) $|\mathcal{S}| = 625.$

Policy Evaluation:

- Batch PG – matrix inversion
- BPMD variants – matrix-vector multiplication

Stochastic Setting – IT

Policy Evaluation with Independent Trajectories



(a) $|\mathcal{S}| = 225.$ **(b)** $|\mathcal{S}| = 400.$ **(c)** $|\mathcal{S}| = 625.$

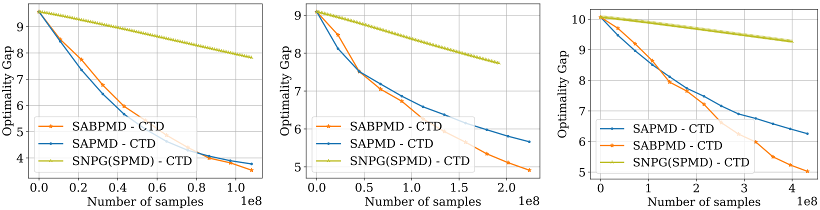
Runtime $\approx$ Number of samples

Policy Evaluation

- Batch PG - $\mathcal{O}(|\mathcal{A}|\,|\mathcal{S}|)$ samples
- BPMD variant - $\mathcal{O}(|\mathcal{A}|)$ samples

Stochastic Setting – CTD

Policy Evaluation with CTD



(a) $|\mathcal{S}| = 225.$ (b) $|\mathcal{S}| = 400.$ (c) $|\mathcal{S}| = 625.$

Runtime $\approx$ Number of samples

Policy Evaluation

- Batch PG - $\mathcal{O}(|\mathcal{A}|\,|\mathcal{S}|)$ samples
- BPMD variant - $\mathcal{O}(|\mathcal{A}|)$ samples

Cheap Per-Iteration Computation

Runtime when Executing the Last Iteration

Method	$\lvert\mathcal{S}\rvert$	Q^{π} Estimation	# (seconds)
SABPMD	400	IT	2.9
SAPMD	400	IT	1192.6
SABPMD	625	IT	2.9
SAPMD	625	IT	1863.5
SABPMD	400	CTD	4.9
SAPMD	400	CTD	1976.5
SABPMD	625	CTD	5.1
SAPMD	625	CTD	3176.5

Conclusion

- BPMD variants with cheaper per-iteration complexity than batch PG methods
- Establish iteration complexities in deterministic/stochastic setting
- Establish sample complexities with two evaluation subroutines

Paper

- Lan, G., Li, Y. and Zhao, T., 2022. Block Policy Mirror Descent. arXiv preprint arXiv:2201.05756

What are still open?

- Dependence of sample complexities on the state space
- PMD (batch PG) + asynchronous policy update