Supplementary Material

C. Proof for CPI, CPI(α), API(α)

We begin by proving the following result:

Theorem 1. *At each iteration $k < k^*$ of CPI (Equation (3)), the expected loss satisfies:*

$$\mu(v_{\pi_*} - v_{\pi_k}) \leq \frac{C^{(1,0)}}{(1-\gamma)^2} \sum_{i=1}^{k} \alpha_i \epsilon_i + e^{\left\{(1-\gamma)\sum_{i=1}^{k} \alpha_i\right\}} V_{\max}.$$

Proof. Using the facts that $T_{\pi_{k+1}} v_{\pi_k} = (1-\alpha_{k+1}) v_{\pi_k} + \alpha_{k+1} T_{\pi_{k+1}} v_{\pi_k}$ and the notation $e_{k+1} = \max_{\pi'} T_{\pi'} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k}$, we have:

$$\begin{aligned}
v_{\pi_*} - v_{\pi_{k+1}} &= v_{\pi_*} - T_{\pi_{k+1}} v_{\pi_k} + T_{\pi_{k+1}} v_{\pi_k} - T_{\pi_{k+1}} v_{\pi_{k+1}} \\
&= v_{\pi_*} - (1-\alpha_{k+1}) v_{\pi_k} - \alpha_{k+1} T_{\pi'_{k+1}} v_{\pi_k} + \gamma P_{\pi_{k+1}} (v_{\pi_k} - v_{\pi_{k+1}}) \\
&= (1-\alpha_{k+1})(v_{\pi_*} - v_{\pi_k}) + \alpha_{k+1}(T_{\pi_*} v_{\pi_*} - T_{\pi_*} v_{\pi_k}) + \alpha_{k+1}(T_{\pi_*} v_{\pi_k} - T_{\pi'_{k+1}} v_{\pi_k}) + \gamma P_{\pi_{k+1}} (v_{\pi_k} - v_{\pi_{k+1}}) \\
&\leq \left[(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}\right] (v_\pi - v_{\pi_k}) + \alpha_{k+1} e_{k+1} + \gamma P_{\pi_{k+1}} (v_{\pi_k} - v_{\pi_{k+1}}). \qquad (12)
\end{aligned}$$

Using the fact that $v_{\pi_{k+1}} = (I - \gamma P_{\pi_{k+1}})^{-1} r$, and the fact that $(I - \gamma P_{\pi_{k+1}})^{-1}$ is non-negative, we can see that

$$\begin{aligned}
v_{\pi_k} - v_{\pi_{k+1}} &= (I - \gamma P_{\pi_{k+1}})^{-1} (v_{\pi_k} - \gamma P_{\pi_{k+1}} v_{\pi_k} - r) \\
&= (I - \gamma P_{\pi_{k+1}})^{-1} (T_{\pi_k} v_{\pi_k} - T_{\pi_{k+1}} v_{\pi_k}) \\
&\leq (I - \gamma P_{\pi_{k+1}})^{-1} \alpha_{k+1} e_{k+1}.
\end{aligned}$$

Putting this back in Equation (12), we obtain:

$$v_{\pi_*} - v_{\pi_{k+1}} \leq \left[(1-\alpha_{k+1}) I + \alpha_{k+1} \gamma P_{\pi_*}\right] (v_\pi - v_{\pi_k}) + \alpha_{k+1} (I - \gamma P_{\pi_{k+1}})^{-1} e_{k+1}.$$

Define the matrix $Q_k = [(1-\alpha_k) I + \alpha_k \gamma P_{\pi_*}]$, the set $\mathcal{N}_{i,k} = \{j; k-i+1 \leq j \leq k\}$ (this set contains exactly i elements), the matrix $R_{i,k} = \prod_{j \in \mathcal{N}_{i,k}} Q_j$, and the coefficients $\beta_k = 1 - \alpha_k(1-\gamma)$ and $\delta_k = \prod_{i=1}^{k} \beta_k$. By repeatedly using the fact that the matrices Q_k are non-negative, we get by induction

$$v_{\pi_*} - v_{\pi_k} \leq \sum_{i=0}^{k-1} R_{i,k} \alpha_{k-i} (I - \gamma P_{\pi_{k-i}})^{-1} e_{k-i} + \delta_k V_{\max}. \qquad (13)$$

Let $\mathcal{P}_j(\mathcal{N}_{i,k})$ be the set of subsets of $\mathcal{N}_{i,k}$ of size j. With this notation we have

$$R_{i,k} = \sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} \zeta_{I,i,k} (\gamma P_{\pi_*})^j$$

where for all subset I of $\mathcal{N}_{i,k}$, we wrote

$$\zeta_{I,i,k} = \left(\prod_{n \in I} \alpha_n\right) \left(\prod_{n \in \mathcal{N}_{i,k} \setminus I} (1-\alpha_n)\right).$$

Therefore, by multiplying Equation (13) by μ, using the definition of the coefficients $c(i)$, and the facts that $\nu \leq (1 -$

$\gamma) d_{\nu,\pi_{k+1}}$, we obtain:

$$\begin{aligned}
\mu(v_{\pi_*} - v_{\pi_k}) &\leq \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{l=0}^{\infty} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} \zeta_{I,i,k} \gamma^{j+l} c(j+l) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\max}. \\
&= \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{l=j}^{\infty} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} \zeta_{I,i,k} \gamma^{l} c(l) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\max} \\
&\leq \frac{1}{1-\gamma} \sum_{i=0}^{k-1} \sum_{j=0}^{i} \sum_{l=0}^{\infty} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} \zeta_{I,i,k} \gamma^{l} c(l) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\max} \\
&= \frac{1}{1-\gamma} \left(\sum_{l=0}^{\infty} \gamma^l c(l) \right) \sum_{i=0}^{k-1} \left(\sum_{j=0}^{i} \sum_{I \in \mathcal{P}_j(\mathcal{N}_{i,k})} \zeta_{I,i,k} \right) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\max} \\
&= \frac{1}{1-\gamma} \left(\sum_{l=0}^{\infty} \gamma^l c(l) \right) \sum_{i=0}^{k-1} \left(\prod_{j \in \mathcal{N}_{i,k}} (1 - \alpha_j + \alpha_j) \right) \alpha_{k-i} \epsilon_{k-i} + \delta_k V_{\max} \\
&= \frac{1}{1-\gamma} \left(\sum_{l=0}^{\infty} \gamma^l c(l) \right) \left(\sum_{i=0}^{k-1} \alpha_{k-i} \epsilon_{k-i} \right) + \delta_k V_{\max}.
\end{aligned}$$

Now, using the fact that for $x \in (0,1)$, $\log(1-x) \leq -x$, we can observe that

$$\log \delta_k = \log \prod_{i=1}^{k} \beta_i = \sum_{i=1}^{k} \log \beta_i = \sum_{i=1}^{k} \log(1 - \alpha_i(1-\gamma)) \leq -(1-\gamma) \sum_{i=1}^{k} \alpha_i.$$

As a consequence, we get $\delta_k \leq e^{-(1-\gamma)\sum_{i=1}^{k} \alpha_i}$. □

In the analysis of CPI, Kakade & Langford (2002) show that the learning steps that ensure the nice performance guarantee of CPI satisfy $\alpha_k \geq \frac{(1-\gamma)\epsilon}{12\gamma V_{\max}}$, the right term $e^{\left\{(1-\gamma)\sum_{i=1}^{k} \alpha_i\right\}}$ above tends 0 exponentially fast, and we get the following corollary that shows that CPI has a performance bound with the coefficient $C^{(1,0)}$ of API in a number of iterations $O\left(\frac{\log \frac{1}{\epsilon}}{\epsilon}\right)$.

Corollary 1. *The smallest (random) iteration $k^\dagger$ such that $\frac{\log \frac{V_{\max}}{\epsilon}}{1-\gamma} \leq \sum_{i=1}^{k^\dagger} \alpha_i \leq \frac{\log \frac{V_{\max}}{\epsilon}}{1-\gamma} + 1$ is such that $k^\dagger \leq \frac{12\gamma V_{\max} \log \frac{V_{\max}}{\epsilon}}{\epsilon(1-\gamma)^2}$ and the policy $\pi_{k^\dagger}$ satisfies:*

$$\mu(v_{\pi_*} - v_{\pi_{k^\dagger}}) \leq \left(\frac{C^{(1,0)} \left(\sum_{i=1}^{k^\dagger} \alpha_i \right)}{(1-\gamma)^2} + 1 \right) \epsilon \leq \left(\frac{C^{(1,0)} \left(\log \frac{V_{\max}}{\epsilon} + 1 \right)}{(1-\gamma)^3} + 1 \right) \epsilon.$$

Since the proof is based on a generalization of the analysis of API and thus does not use any of the specific properties of CPI, it turns out that the results we have just given can straightforwardly be specialized to CPI(α).

Corollary 2. *Assume we run CPI(α) for some $\alpha \in (0,1)$, that is CPI (Equation (3)) with $\alpha_k = \alpha$ for all k.*

$$\textit{If } k = \left\lceil \frac{\log \frac{V_{\max}}{\epsilon}}{\alpha(1-\gamma)} \right\rceil, \quad \textit{then } \mu(v_{\pi_*} - v_{\pi_k}) \leq \frac{\alpha(k+1)C^{(1,0)}}{(1-\gamma)^2} \epsilon \leq \left(\frac{C^{(1,0)} \left(\log \frac{V_{\max}}{\epsilon} + 1 \right)}{(1-\gamma)^3} + 1 \right) \epsilon.$$

The above bound for CPI(α) involves the factor $\frac{1}{(1-\gamma)^3}$. A precise examination of the proof shows that this amplification is due to the fact that the approximate greedy operator uses the distribution $d_{\pi_k,\nu} \geq (1-\gamma)\nu$ instead of ν (for API). In fact, using a very similar proof, it is easy to show that API(α) satisfies the following result.

Corollary 3. *Assume API(α) is run for some* $\alpha \in (0,1)$.

$$\textit{If } k = \left\lceil \frac{\log \frac{V_{\max}}{\epsilon}}{\alpha(1-\gamma)} \right\rceil, \quad \textit{then } \mu(v_{\pi_*} - v_{\pi_k}) \leq \frac{\alpha(k+1)C^{(1,0)}}{(1-\gamma)}\epsilon \leq \left(\frac{C^{(1,0)}\left(\log \frac{V_{\max}}{\epsilon} + 1\right)}{(1-\gamma)^2} + 1 \right)\epsilon.$$

D. More details on the Numerical Simulations

Domain and Approximations In our experiments, a Garnet is parameterized by 4 parameters and is written $G(n_S, n_A, b, p)$: n_S is the number of states, n_A is the number of actions, b is a branching factor specifying how many possible next states are possible for each state-action pair (b states are chosen uniformly at random and transition probabilities are set by sampling uniform random $b-1$ cut points between 0 and 1) and p is the number of features (for linear function approximation). The reward is state-dependent: for a given randomly generated Garnet problem, the reward for each state is uniformly sampled between 0 and 1. Features are chosen randomly: Φ is a $n_S \times p$ feature matrix of which each component is randomly and uniformly sampled between 0 and 1. The discount factor γ is set to 0.99 in all experiments.

All the algorithms we have discussed in the paper need to repeatedly compute $\mathcal{G}_\epsilon(\rho, v)$ for some distribution $\rho = \nu$ or $\rho = d_{\pi,\nu}$. In other words, they must be able to make calls to an approximate greedy operator applied to the value v of some policy for some distribution ρ. To implement this operator, we compute a noisy estimate of the value v with a uniform white noise $u(\iota)$ of amplitude ι, then projects this estimate onto the space spanned by Φ with respect to the ρ-quadratic norm (projection that we write $\Pi_{\Phi,\rho}$), and then applies the (exact) greedy operator on this projected estimate. In a nutshell, one call to the approximate greedy operator $\mathcal{G}_\epsilon(\rho, v)$ amounts to compute $\mathcal{G}\Pi_{\Phi,\rho}(v + u(\iota))$.

Simulations We have run series of experiments, in which we callibrated the perturbations (noise, approximations) so that the algorithm are significantly perturbed but no too much (we do not want their behavior to become too erratic). After trial and error, we ended up considering the following setting. We used Garnet problems $G(n_S, n_A, b, p)$ with the number of states $n_S \in \{50, 100, 200\}$, the number of actions $n_A \in \{2, 5, 10\}$, the branching factor $b \in \{1, 2, 10\}\}$ ($b = 1$ corresponds to deterministic problems), the number of features to approximate the value $p = \frac{n_S}{10}$, and the noise level $\iota = 0.1$ (10%).

In addition to Figure 2 that shows the statistics overall for the all the parameter instances, Figure 3, 4 and 5 display statistics that are respectively conditioned on the values of n_S, n_A and b, which gives some insight on the influence of these parameters.

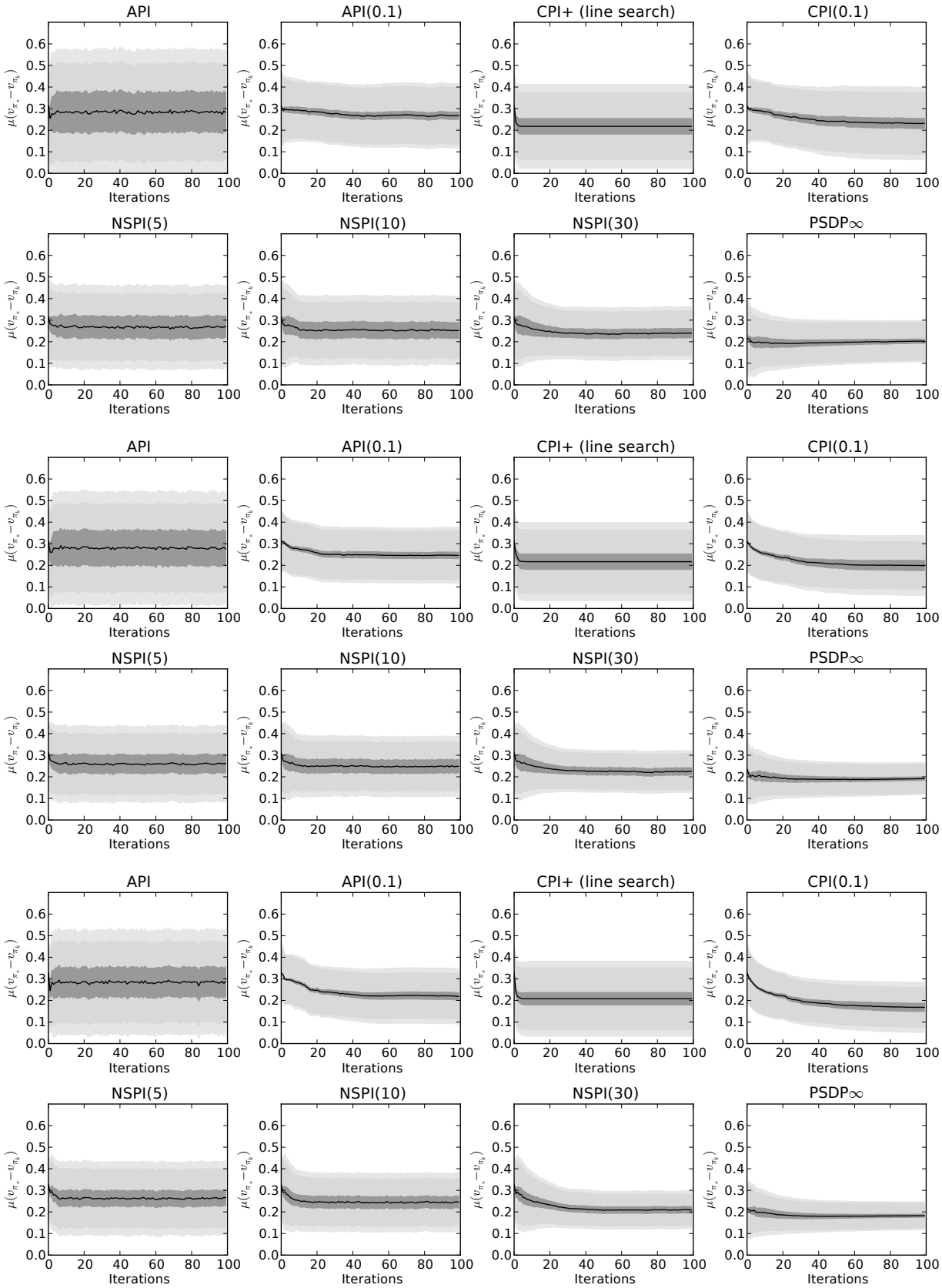

Figure 3. Statistics conditioned on the number of states. Top: $n_S = 50$. Middle: $n_S = 100$. Bottom $n_S = 200$.

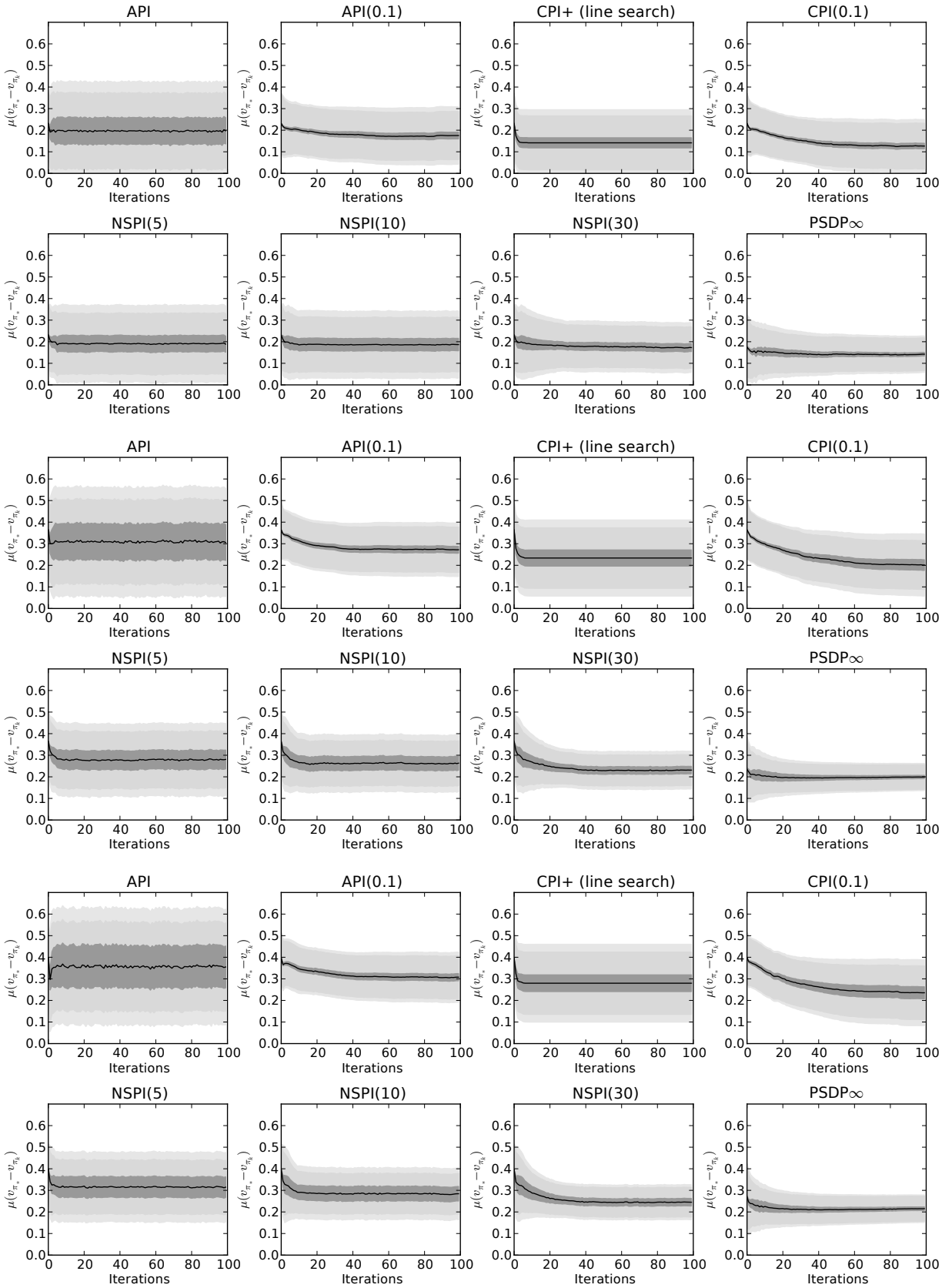

Figure 4. Statistics conditioned on the number of actions. Top: $n_A = 2$. Middle: $n_A = 5$. Bottom $n_a = 10$.

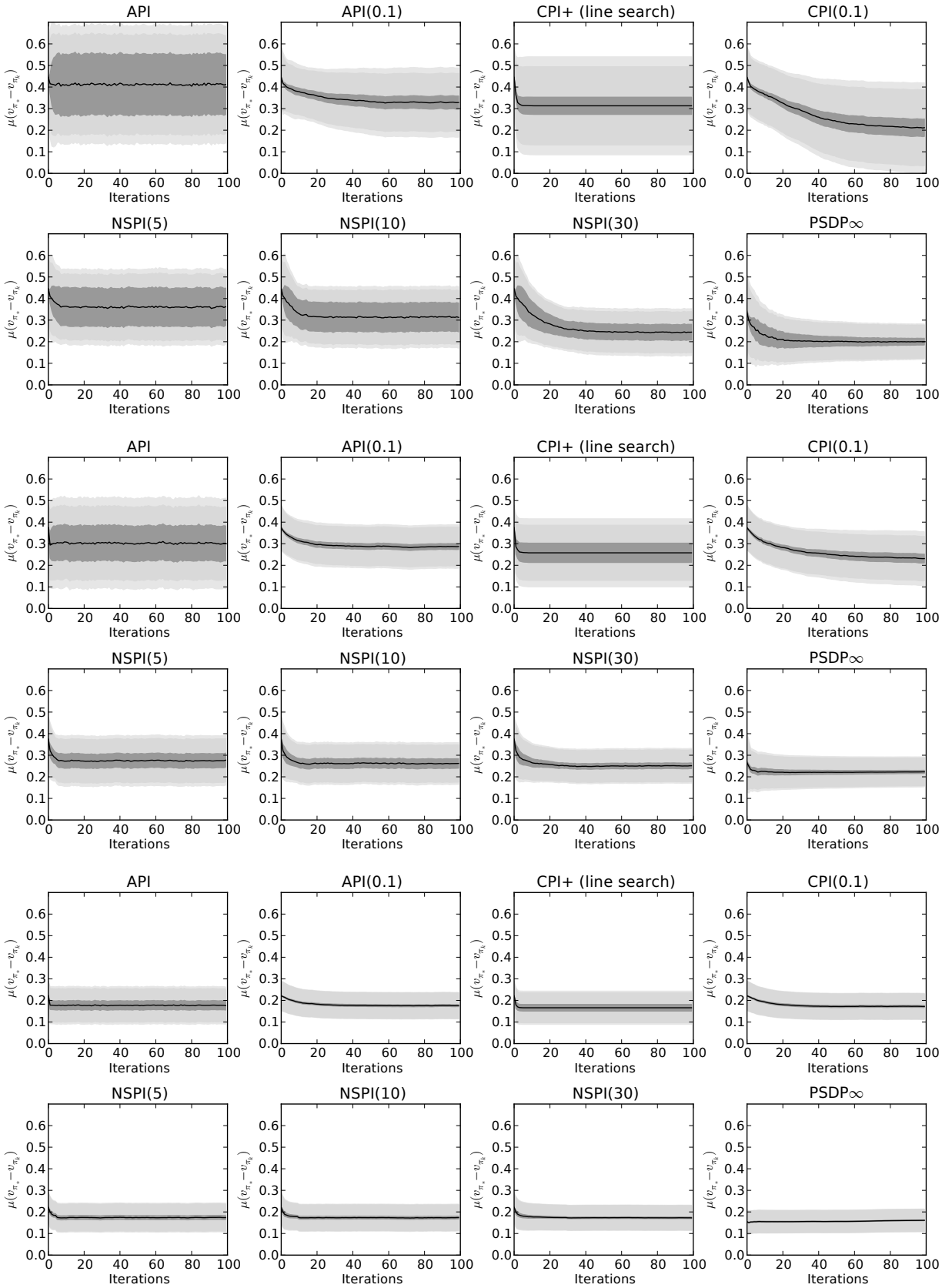

Figure 5. Statistics conditioned on the branching factor. Top: $b = 1$ (deterministic). Middle: $b = 2$. Bottom $b = 10$.