

Summary

- ✓ We propose a model-free algorithm called ASPIC that smoothes the cost function by applying an **inf-convolution** aiming to speedup convergence of policy optimization
- ✓ ASPIC bridges two well-known policy optimization methods: **direct cost optimization** and the **cross-entropy method**
- ✓ We show that **intermediate levels of smoothing are optimal**

Path Integral Control

The optimization objective: **entropy-regularized expected cost**

$$C(p_u) = \langle V(\tau) \rangle_{p_u} + \gamma KL(p_u || p_0) \quad \text{with } V(\tau) = \int_0^T V(x_t, t) dt$$

$V(\tau)$: state cost of trajectory $\tau = \{x_t | \forall t : 0 < t \leq T\}$

$p_0(\tau)$: **uncontrolled dynamics**

$p_u(\tau)$: induced probability over **controlled state trajectories**

Parameter γ controls the strength of the regularization $KL(p_u || p_0)$

(Approach I) → Direct-Cost Optimization

Gradient descent over parametrized path probabilities p_{u_θ}

$$\nabla_\theta C(p_{u_\theta}) = \left\langle S_{p_{u_\theta}}^\gamma(\tau) \nabla_\theta \log p_{u_\theta}(\tau) \right\rangle_{p_{u_\theta}}$$

with the stochastic cost $S_{p_{u_\theta}}^\gamma(\tau) := V(\tau) + \gamma \log \frac{p_{u_\theta}(\tau)}{p_0(\tau)}$

(Approach II) → Cross-Entropy Method

KL divergence minimization has analytical solution

$$p^* := \arg \min_p C(p) = \frac{1}{Z} p_0(\tau) \exp \left(-\frac{1}{\gamma} V(\tau) \right), \quad \text{with } Z = \left\langle \exp \left(-\frac{1}{\gamma} V(\tau) \right) \right\rangle_{p_0}$$

The Cross-Entropy method results in the gradient

$$\nabla_\theta KL(p^* || p_{u_\theta}) = \frac{1}{Z_{p_{u_\theta}}} \left\langle \exp \left(-\frac{1}{\gamma} S_{p_{u_\theta}}^\gamma(\tau) \right) \nabla_\theta \log p_{u_\theta}(\tau) \right\rangle_{p_{u_\theta}}$$

Approach I and Approach II share the same global minima
Which approach is better?

Smoothing Stochastic Control Problems

Inf-Convolution applied the original cost

- ✓ The smoothed cost $J^\alpha(\theta)$ (also known as **Moreau envelope** function)

$$J^\alpha(\theta) = \inf_{\theta'} \alpha KL(p_{u_{\theta'}} || p_{u_\theta}) + \underbrace{\gamma KL(p_{u_{\theta'}} || p_0) + \langle V(\tau) \rangle_{p_{u_{\theta'}}}}_{C(p_{u_{\theta'}})}$$

- Parameter α penalizes the deviation of $p_{u_{\theta'}}$ from p_{u_θ}
- Parameter γ penalizes the deviation of $p_{u_{\theta'}}$ from the uncontrolled dynamics p_0

- ✓ **Preserves global minima**

Smoothing bridges Direct-Cost Optimization and the Cross-Entropy Method

Smoothing has analytical solution for path integral control

The smoothed cost becomes

$$J^\alpha(\theta) = -(\gamma + \alpha) \log \left\langle \exp \left(-\frac{1}{\gamma + \alpha} S_{p_{u_\theta}}^\gamma(\tau) \right) \right\rangle_{p_{u_\theta}}$$

and its gradient

$$\nabla_\theta J^\alpha(\theta) = -\frac{\alpha}{Z_{p_{u_\theta}}} \left\langle \exp \left(-\frac{1}{\gamma + \alpha} S_{p_{u_\theta}}^\gamma(\tau) \right) \nabla_\theta \log p_{u_\theta}(\tau) \right\rangle_{p_{u_\theta}}$$

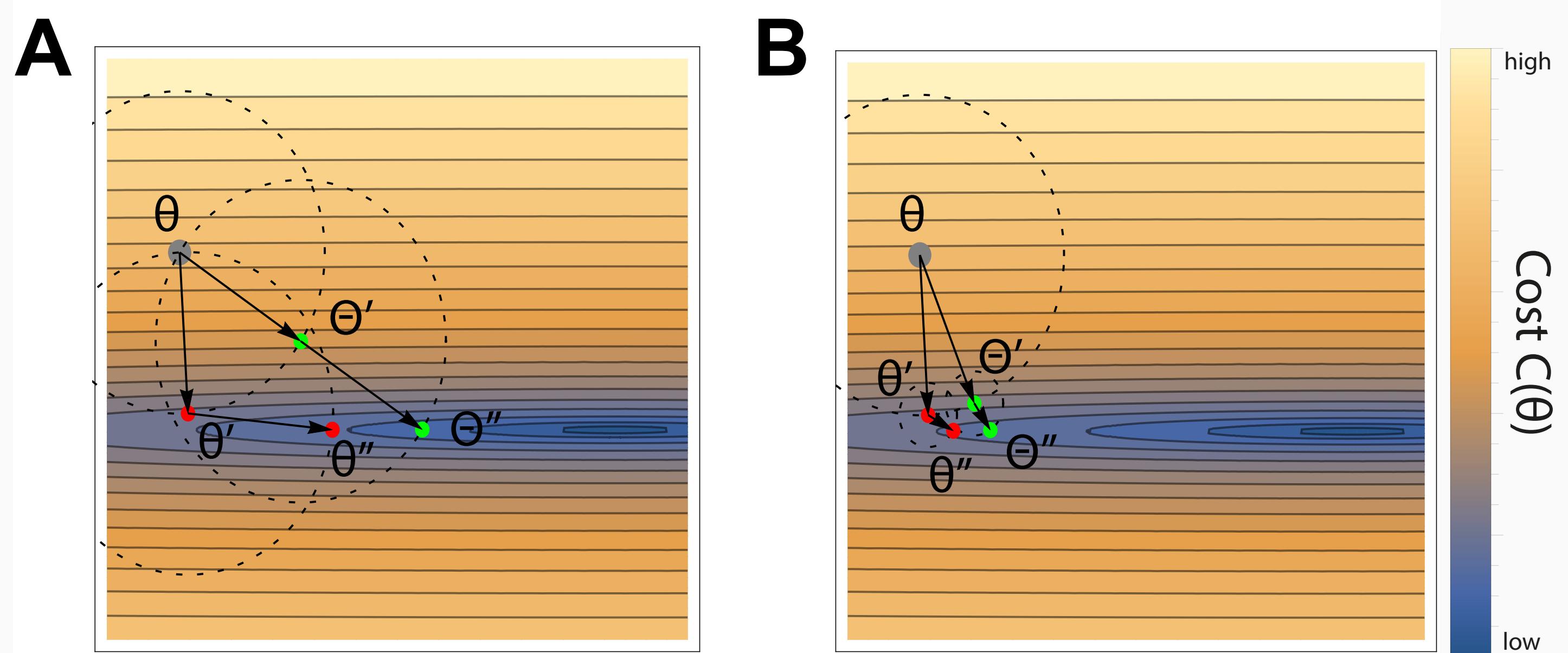
$\alpha \rightarrow \infty$ (weak smoothing) recovers **Direct-Cost Optimization**

$\alpha \rightarrow 0$ (strong smoothing) recovers **Cross-Entropy Method**

Smoothing Accelerates Policy Optimization

A two-step task with 2D cost landscape $C(\theta)$ parametrized by θ

Theorem: a smoothed update $\Theta' = \Theta_{\mathcal{E}}^{J^\alpha}(\theta)$ with stepsize \mathcal{E} followed by a direct update $\Theta'' = \Theta_{\mathcal{E}'}^C(\Theta')$ with stepsize \mathcal{E}' is optimal, for all \mathcal{E}, α .



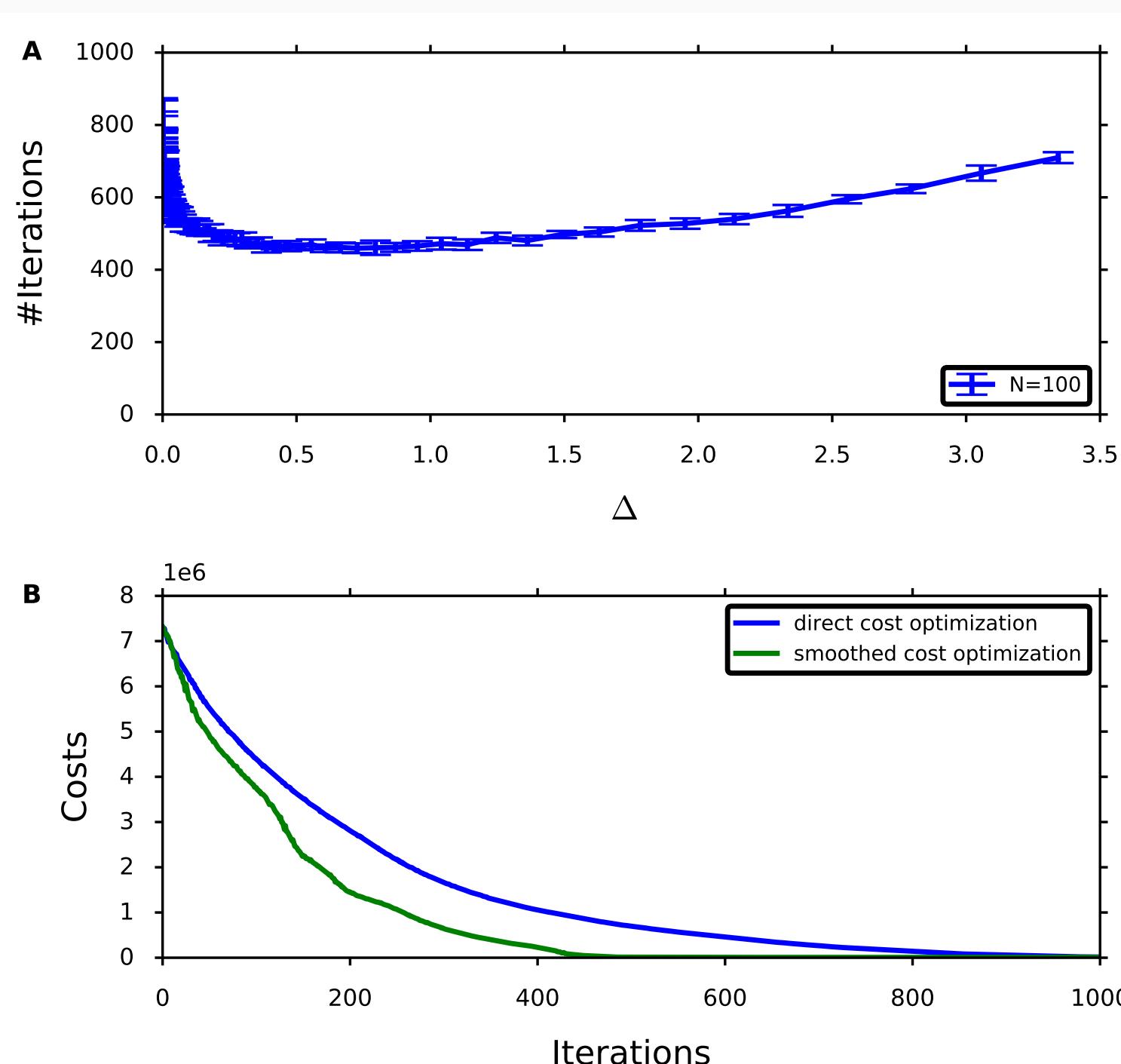
✓ (B) When the size of the second step becomes small $\mathcal{E}' \ll \mathcal{E}$, the smoothed update $\theta \rightarrow \Theta'$ becomes more similar to the direct update $\theta \rightarrow \theta'$.

The ASPIC Algorithm

- ▷ **Adapts the smoothing parameter** α to keep the variance of the gradient estimator at a predefined level, independently of the number of samples N per iteration
- ▷ **Performs natural gradient updates** $\theta_{n+1} = \theta_n - \beta^{-1} F^{-1} \nabla_\theta J^\alpha(\theta')|_{\theta'=\theta_n}$
- ✓ Two hyper-parameters:
 - The smoothing strength Δ (a bound on the variance)
 - The size of the trust regions \mathcal{E}

Experiments

▷ Linear-Quadratic control problems (“exact” policies)

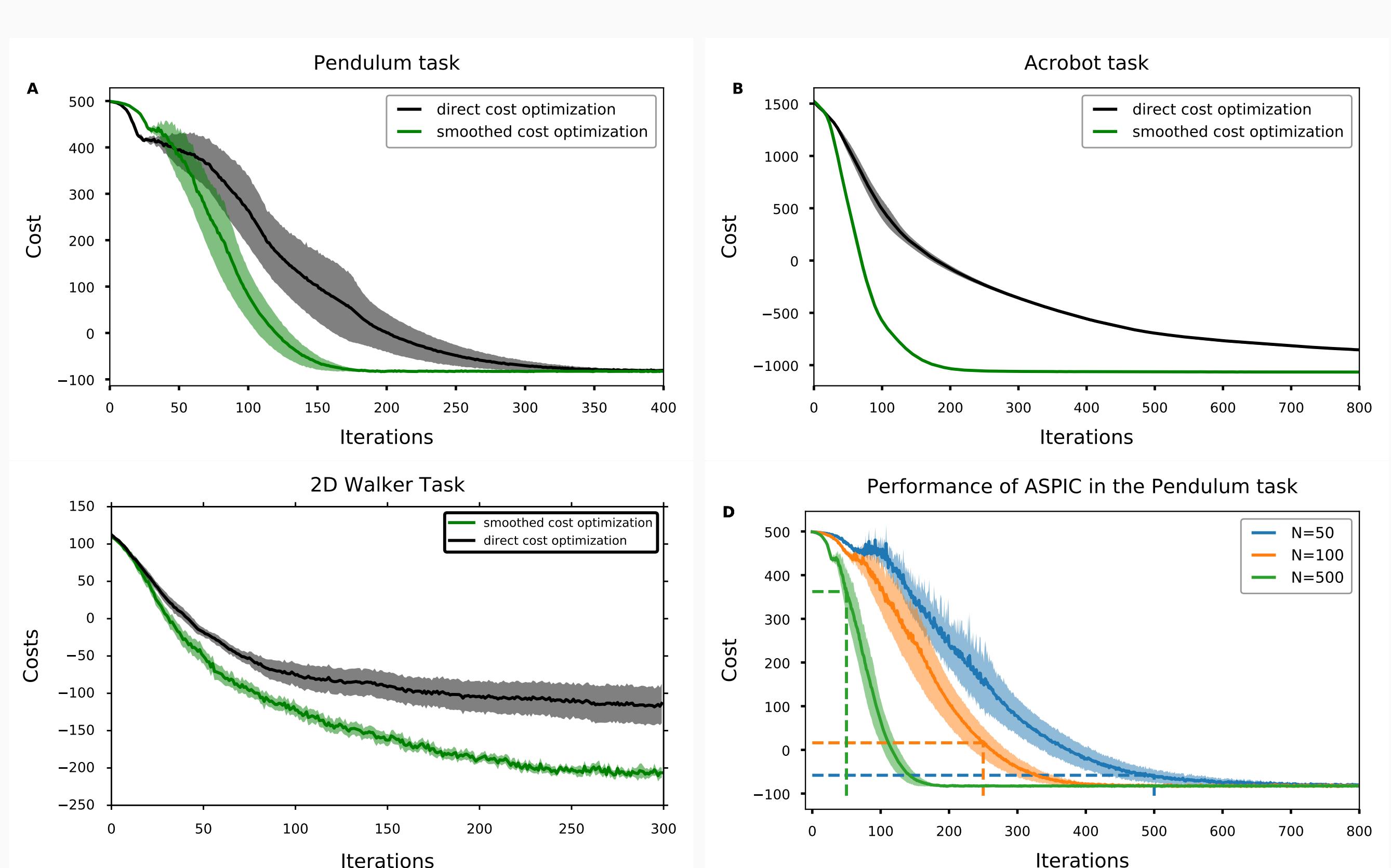


✓ Smoothing ($\Delta > 0$) accelerates

✓ For too large Δ : gradient estimation is not reliable

✓ Optimal smoothing at intermediate values of $\Delta \approx 0.4$

▷ Nonlinear Control Problems (“approximated” policies)



✓ A strong performance boost can also be achieved

✓ Scales well to complex problems using neural network policies