Learning Rate Adaptation by Line Search in Evolution Strategies with Recombination

Armand Gissler, Anne Auger, Nikolaus Hansen

INRIA & CMAP, École Polytechnique, Palaiseau, France

GECCO 2022

The $(\mu/\mu, \lambda)$-ES

Goal:

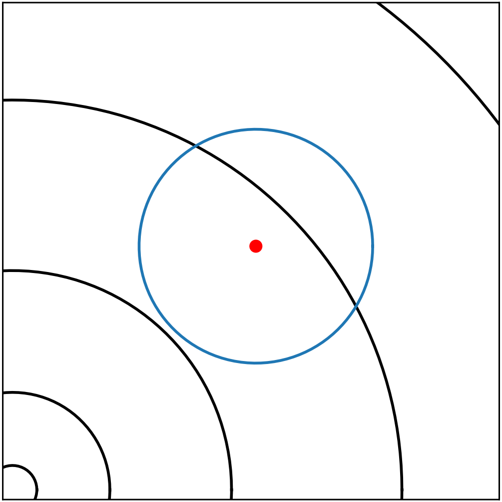
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x).$$

Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

-
-
-
-

The $(\mu/\mu, \lambda)$-ES



The $(\mu/\mu, \lambda)$-ES

Goal:

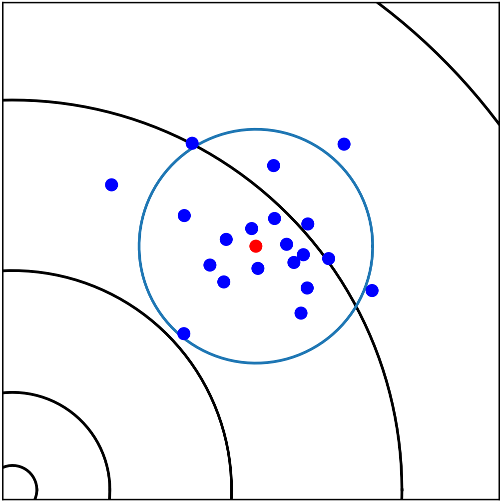
$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x).$$

Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

- **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
-
-
-

The $(\mu/\mu, \lambda)$-ES



The $(\mu/\mu, \lambda)$-ES

Goal:

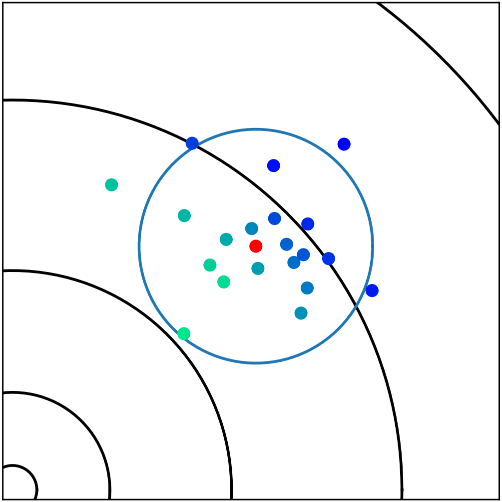
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x).$$

Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

- **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
- **Sort** $f(X_t + \sigma_t U_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(X_t + \sigma_t U_{t+1}^{\lambda:\lambda})$;
-
-

The $(\mu/\mu, \lambda)$-ES



The $(\mu/\mu, \lambda)$-ES

Goal:

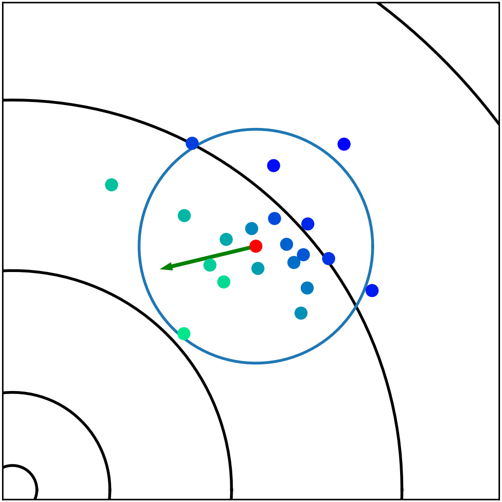
$$\underset{x \in \mathbb{R}^n}{\text{minimize}}\ f(x).$$

Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

- **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
- **Sort** $f(X_t + \sigma_t U_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(X_t + \sigma_t U_{t+1}^{\lambda:\lambda})$;
- **Update the mean** $X_{t+1} = X_t + \kappa \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda}$;
-

The $(\mu/\mu, \lambda)$-ES



The $(\mu/\mu, \lambda)$-ES

Goal:

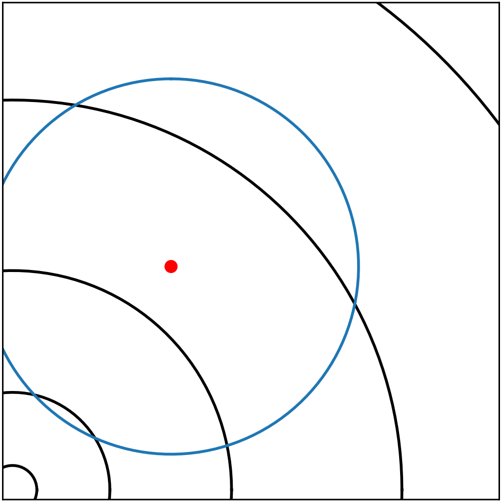
$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \; f(x).$$

Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

- **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
- **Sort** $f(X_t + \sigma_t U_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(X_t + \sigma_t U_{t+1}^{\lambda:\lambda})$;
- **Update the mean** $X_{t+1} = X_t + \kappa \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda}$;
- **Update the step-size** $\sigma_{t+1} = \bar{\sigma}(X_{t+1}, (U_{t+1}^{i:\lambda})_i, \sigma_t)$.

The $(\mu/\mu, \lambda)$-ES



The $(\mu/\mu, \lambda)$-ES

Goal:

$$\underset{x \in \mathbb{R}^n}{\text{minimize}} \ f(x).$$

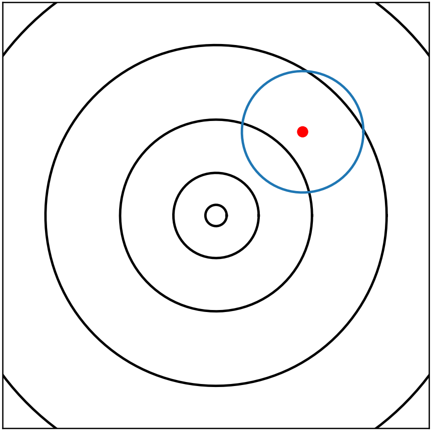
Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

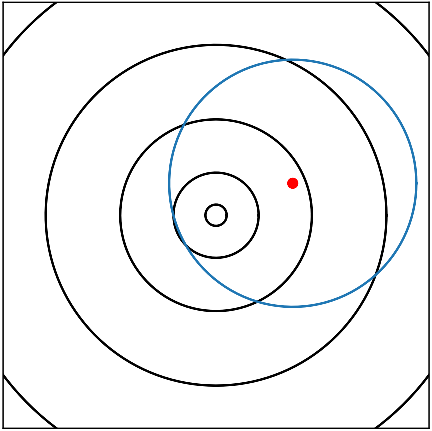
- **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
- **Sort** $f(X_t + \sigma_t U_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(X_t + \sigma_t U_{t+1}^{\lambda:\lambda})$;
- **Update the mean** $X_{t+1} = X_t + \kappa \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda}$;
- **Update the step-size** $\sigma_{t+1} = \bar{\sigma}(X_{t+1}, (U_{t+1}^{i:\lambda})_i, \sigma_t)$.

The coefficient κ is the *learning rate* (of the mean). Usually $\kappa = 1$.

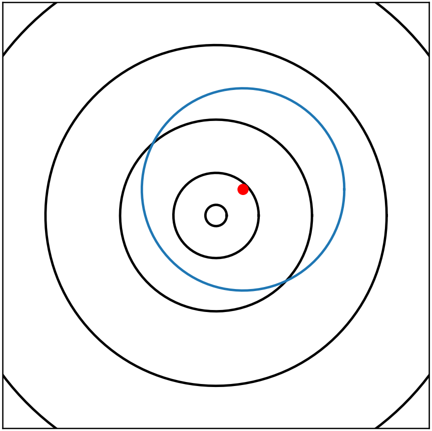
Convergence



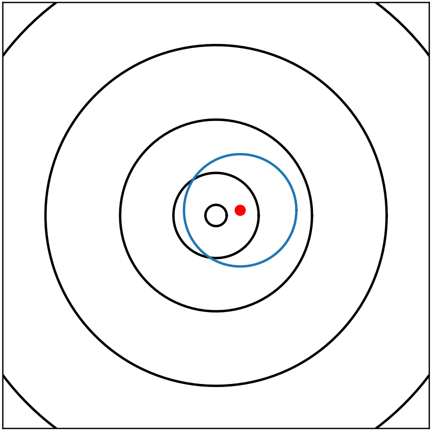
Convergence



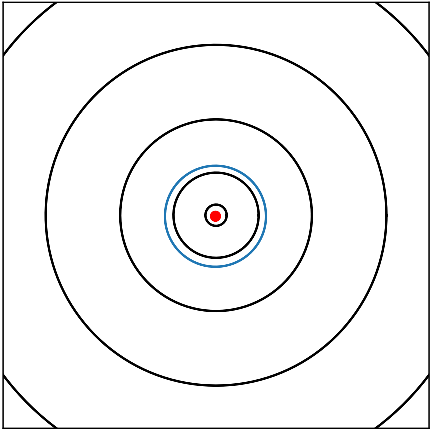
Convergence



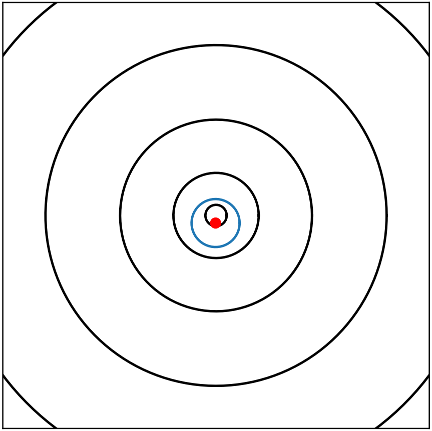
Convergence



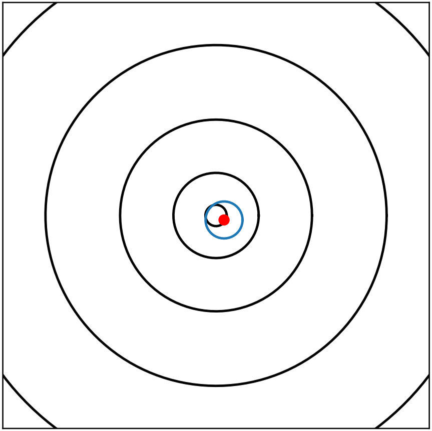
Convergence



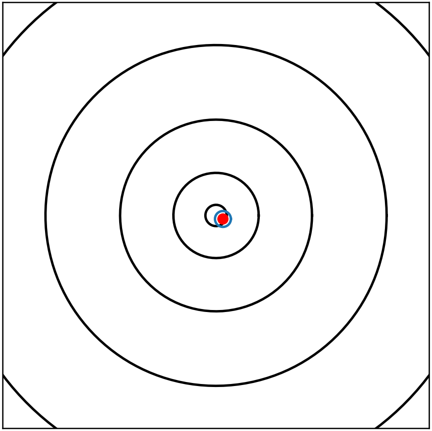
Convergence



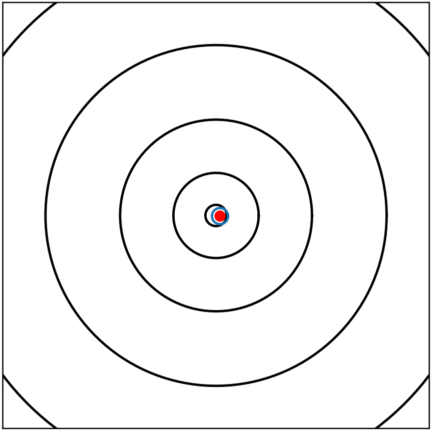
Convergence



Convergence



Convergence

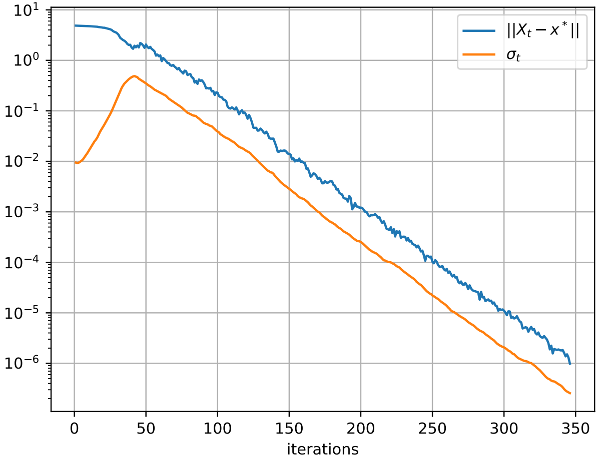


Linear convergence

$$\lim_{t\to\infty} -\frac{1}{\lambda} \ln \frac{\|X_t - x^*\|}{\|X_0 - x^*\|} = \lim_{t\to\infty} -\frac{1}{\lambda} \ln \frac{\sigma_t}{\sigma_0} = \mathrm{CR}.$$

Linear convergence

$$\lim_{t\to\infty} -\frac{1}{\lambda} \ln \frac{\|X_t - x^*\|}{\|X_0 - x^*\|} = \lim_{t\to\infty} -\frac{1}{\lambda} \ln \frac{\sigma_t}{\sigma_0} = \mathrm{CR}.$$



Influence of the learning rate κ on the convergence rate

The convergence rates (on the sphere function) of the $(\mu/\mu, \lambda)$-ES writes as

$$\mathrm{CR} = -\frac{1}{\lambda}\mathbb{E} \ln \left\| X_t + \kappa \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda} \right\| .$$

Influence of the learning rate κ on the convergence rate

The convergence rates (on the sphere function) of the $(\mu/\mu, \lambda)$-ES writes as

$$\mathrm{CR} = -\frac{1}{\lambda}\mathbb{E} \ln \left\| X_t + \kappa\sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda} \right\| .$$

- It does not depend on the starting point, i.e. we can choose $X_t = e_1$;

Influence of the learning rate κ on the convergence rate

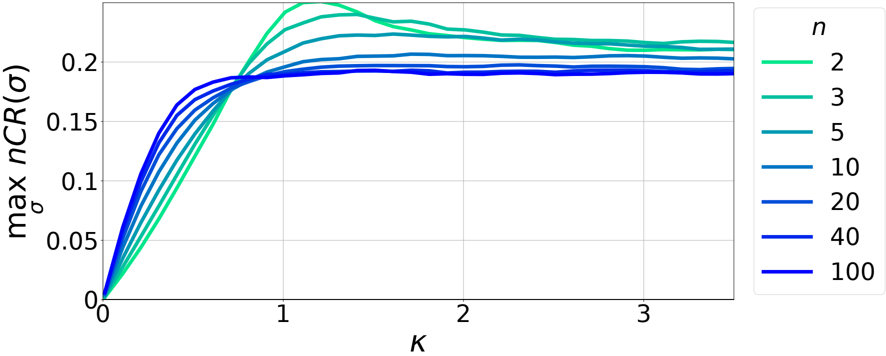
The convergence rates (on the sphere function) of the $(\mu/\mu, \lambda)$-ES writes as

$$\mathrm{CR} = -\frac{1}{\lambda}\mathbb{E}\ln\left\|X_t + \kappa\sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda}\right\|.$$

- It does not depend on the starting point, i.e. we can choose $X_t = e_1$;
- We suppose here that the step-size is proportional to the distance to the optimum $\sigma_t = \alpha\|X_t - x^*\|$.

Influence of the learning rate κ on the convergence rate

$$\mathrm{CR} = -\frac{1}{\lambda}\mathbb{E}\ln\left\| X_t + \kappa\sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda} \right\| .$$



The $(\mu/\mu, \lambda)$-ES with dynamic learning rate

Algorithm One $(\mu/\mu, \lambda)$-ES iteration

Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$

- **Sample** $U_{t+1}^1, \ldots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
- **Sort** $f(X_t + \sigma_t U_{t+1}^{1:\lambda}) \leqslant \cdots \leqslant f(X_t + \sigma_t U_{t+1}^{\lambda:\lambda})$;
- **Update the mean** $X_{t+1} = X_t + \kappa \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{\mu:\lambda}$;
- **Update the step-size** $\sigma_{t+1} = \bar{\sigma}(X_{t+1}, (U_{t+1}^{i:\lambda})_i, \sigma_t)$.

The $(\mu/\mu, \lambda)$-ES with dynamic learning rate

Algorithm One $(\mu/\mu, \lambda)$-ES iteration with dynamic learning rate

---

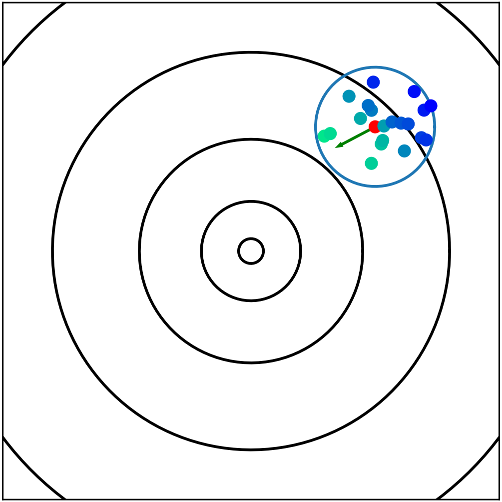
Given: $X_t \in \mathbb{R}^n$, $\sigma_t > 0$, $\kappa_t > 0$

- **Sample** $U_{t+1}^1, \dots, U_{t+1}^\lambda \sim \mathcal{N}(0, I_n)$ i.i.d. ;
- **Sort** $f(X_t + \sigma_t U_{t+1}^{1:\lambda}) \leqslant \dots \leqslant f(X_t + \sigma_t U_{t+1}^{\lambda:\lambda})$;
- **Compute the learning rate** $\kappa_{t+1} = \bar{\kappa}(X_t, \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda}, \kappa_t)$;
- **Update the mean** $X_{t+1} = X_t + \kappa_{t+1} \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda}$;
- **Update the step-size** $\sigma_{t+1} = \bar{\sigma}(X_{t+1}, (U_{t+1}^{i:\lambda})_i, \sigma_t)$.

Examples

- Fixed learning rate $\bar{\kappa}(x, v, \kappa) = 1$,

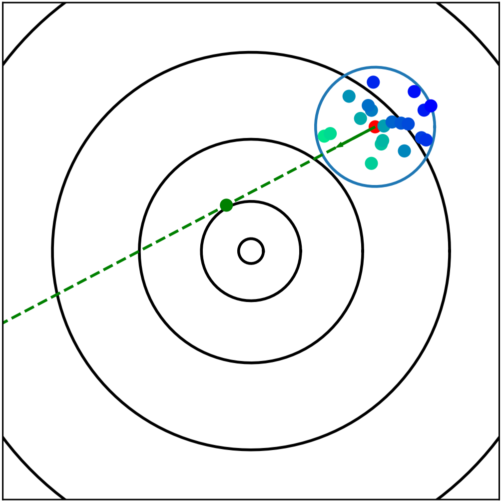
Examples



Examples

- Fixed learning rate $\bar{\kappa}(x, v, \kappa) = 1$,
- *Perfect line search* $\bar{\kappa}(x, v, \kappa) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$.

Examples



Convergence results

Suppose the learning rate update $\bar{\kappa}$ is independent of the parameter κ, and satisfies

Convergence results

Suppose the learning rate update $\bar{\kappa}$ is independent of the parameter κ, and satisfies

(A1) Scaling-invariance:

$$\bar{\kappa}(rx, rv) = \bar{\kappa}(x, v).$$

Convergence results

Suppose the learning rate update $\bar{\kappa}$ is independent of the parameter κ, and satisfies

(A1) Scaling-invariance:

$$\bar{\kappa}(rx, rv) = \bar{\kappa}(x, v).$$

(A2) Rotation-invariance:

$$\bar{\kappa}(Rx, Rv) = \bar{\kappa}(x, v).$$

Convergence results

Suppose the learning rate update $\bar{\kappa}$ is independent of the parameter κ, and satisfies

(A1) Scaling-invariance:

$$\bar{\kappa}(rx, rv) = \bar{\kappa}(x, v).$$

(A2) Rotation-invariance:

$$\bar{\kappa}(Rx, Rv) = \bar{\kappa}(x, v).$$

Theorem

If $\bar{\kappa} \colon \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}_+$ satisfies (A1) and (A2), then linear convergence holds for $(\mu/\mu, \lambda)$-ES with dynamic learning rate $\bar{\kappa}$, and

$$\mathrm{CR} = -\frac{1}{\lambda + C} \mathbb{E} \left\| X_t + \kappa_{t+1} \sigma_t \sum_{i=1}^{\mu} w_i U_{t+1}^{i:\lambda} \right\| .$$

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.
For a fixed learning rate we know that

$$\lim_{n \to \infty} nCR = -\frac{2\alpha \mathbb{E}\left[\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right] + \alpha^2/\mu_w}{2\lambda},$$

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.
For a fixed learning rate we know that

$$\lim_{n \to \infty} nCR = -\frac{2\alpha \mathbb{E}\left[\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right] + \alpha^2/\mu_w}{2\lambda},$$

given that, for all dimensions n, $\sigma_t = \alpha \|X_t - x^*\|/n$.

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.
For a fixed learning rate we know that

$$\lim_{n \to \infty} nCR = -\frac{2\alpha \mathbb{E}\left[\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right] + \alpha^2/\mu_w}{2\lambda},$$

given that, for all dimensions n, $\sigma_t = \alpha \|X_t - x^*\|/n$.

Theorem

Suppose that $f = \|\cdot\|^2$,

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.
For a fixed learning rate we know that

$$\lim_{n \to \infty} nCR = -\frac{2\alpha \mathbb{E}\left[\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right] + \alpha^2/\mu_w}{2\lambda},$$

given that, for all dimensions n, $\sigma_t = \alpha \|X_t - x^*\|/n$.

Theorem

Suppose that $f = \|\cdot\|^2$, *that* $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$ *(cost-free perfect line search),*

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.
For a fixed learning rate we know that

$$\lim_{n \to \infty} nCR = -\frac{2\alpha \mathbb{E}\left[\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right] + \alpha^2/\mu_w}{2\lambda},$$

given that, for all dimensions n, $\sigma_t = \alpha \|X_t - x^*\|/n$.

Theorem

Suppose that $f = \|\cdot\|^2$, that $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$ (cost-free perfect line search), and that $\sigma_t = \alpha \|X_t - x^*\|/n$.

Asymptotic limit of the convergence rates

This result apply to perfect line search $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$, and to a learning rate fixed to 1 $\bar{\kappa}(x, v) = 1$.
For a fixed learning rate we know that

$$\lim_{n \to \infty} nCR = -\frac{2\alpha \mathbb{E}\left[\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right] + \alpha^2/\mu_w}{2\lambda},$$

given that, for all dimensions n, $\sigma_t = \alpha \|X_t - x^*\|/n$.

Theorem

Suppose that $f = \|\cdot\|^2$, that $\bar{\kappa}(x, v) = \arg\min_{\kappa \geqslant 0} f(x + \kappa v)$ (cost-free perfect line search), and that $\sigma_t = \alpha \|X_t - x^*\|/n$. Then

$$\lim_{n \to \infty} nCR = \frac{\mu_w}{2\lambda} \mathbb{E}\left[\left(\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda}\right)^2 \mathbf{1}_{\sum_{i=1}^{\mu} w_i \mathcal{N}^{i:\lambda} < 0}\right].$$

Numerical estimation of the convergence rates

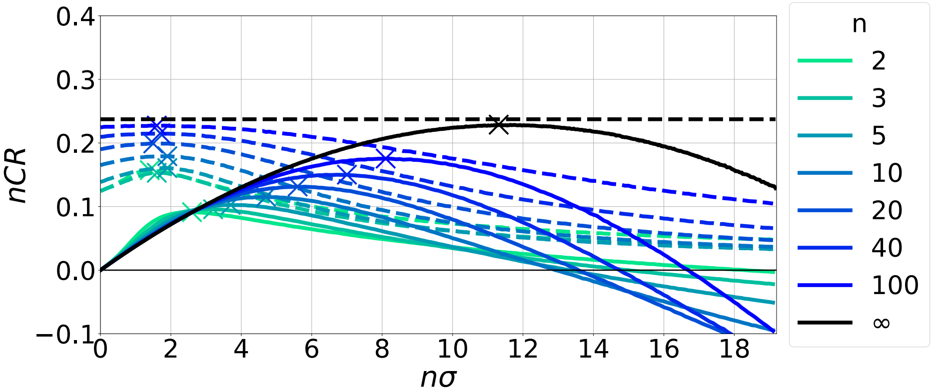
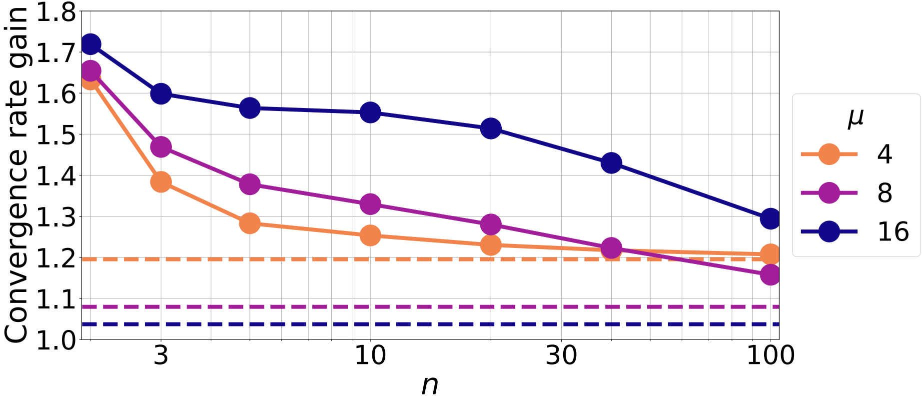


Figure: Convergence rate versus step-size without line search (solid lines) and with perfect line search (dashed lines).

Convergence rate gain due to perfect line search

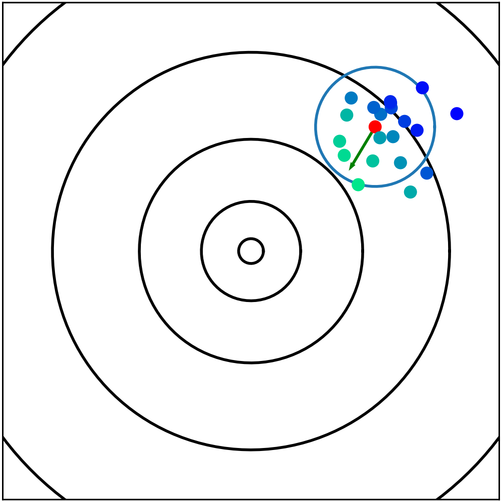


Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

Example: a dichotomic line search



Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

- $\kappa_{0,0} = \kappa_0/2$, $\kappa_{1,0} = \kappa_0 \times 2$
-
    -
    -
    -
    -

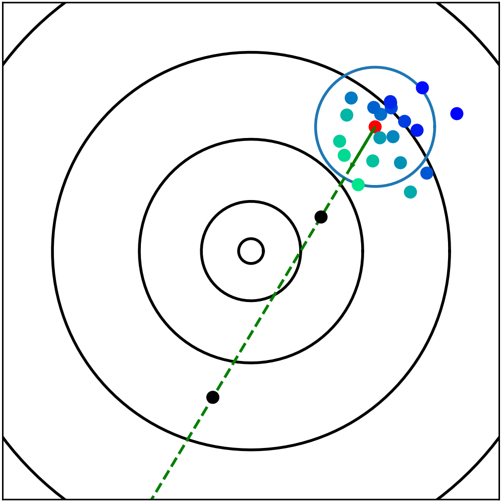
Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

- $\kappa_{0,0} = \kappa_0/2$, $\kappa_{1,0} = \kappa_0 \times 2$
- For $i = 0$:
  - $Y_{\delta,i} = x + \kappa_{\delta,i} v$ for $\delta = 0, 1$
  -
  -
  -

Example: a dichotomic line search



Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

- $\kappa_{0,0} = \kappa_0/2$, $\kappa_{1,0} = \kappa_0 \times 2$
- For $i = 0$:
  - $Y_{\delta,i} = x + \kappa_{\delta,i} v$ for $\delta = 0, 1$
  - $\delta^* = \arg\min_\delta f(Y_{\delta,i})$
  -
  -

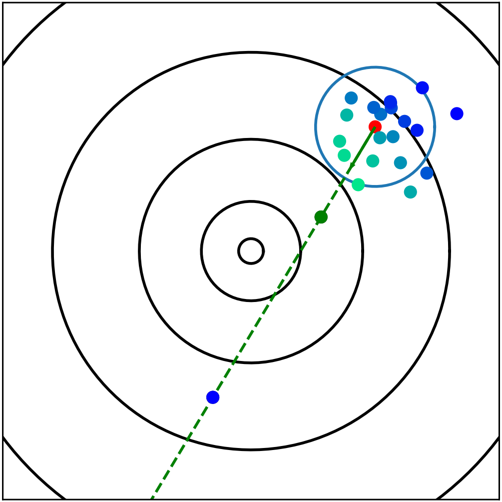
Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

- $\kappa_{0,0} = \kappa_0/2$, $\kappa_{1,0} = \kappa_0 \times 2$
- For $i = 0$:
  - $Y_{\delta,i} = x + \kappa_{\delta,i} v$ for $\delta = 0, 1$
  - $\delta^* = \arg\min_\delta f(Y_{\delta,i})$
  - $\kappa_{\delta^*,i+1} = \kappa_{\delta^*,i}$
  -

Example: a dichotomic line search



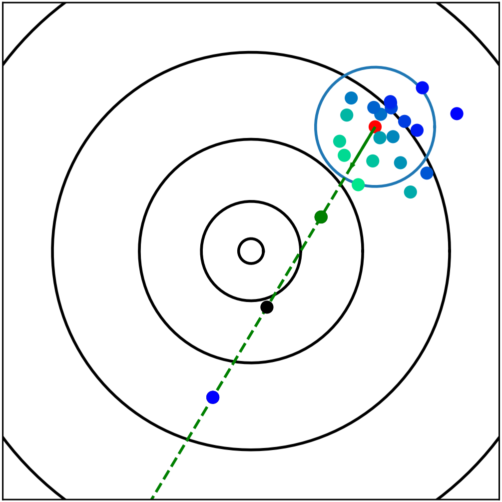
Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

- $\kappa_{0,0} = \kappa_0/2$, $\kappa_{1,0} = \kappa_0 \times 2$
- For $i = 0$:
  - $Y_{\delta,i} = x + \kappa_{\delta,i} v$ for $\delta = 0, 1$
  - $\delta^* = \arg\min_\delta f(Y_{\delta,i})$
  - $\kappa_{\delta^*,i+1} = \kappa_{\delta^*,i}$
  - $\kappa_{1-\delta^*,i+1} = \mathrm{Mean}(\kappa_{\delta^*,i}, \kappa_{1-\delta^*,i})$

Example: a dichotomic line search



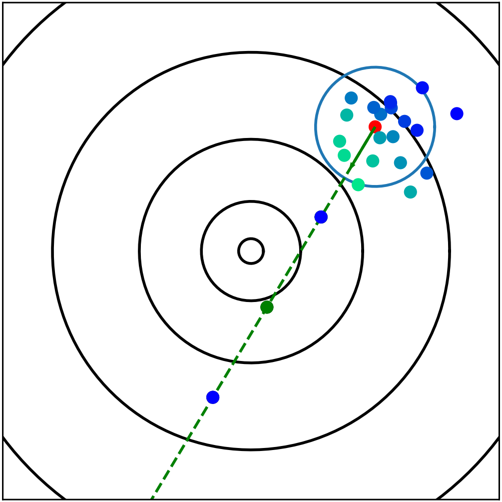
Example: a dichotomic line search

Algorithm Dichotomic line search

Given: $X \in \mathbb{R}^n$, $v \in \mathbb{R}^n$, $\kappa_0 > 0$

- $\kappa_{0,0} = \kappa_0/2$, $\kappa_{1,0} = \kappa_0 \times 2$
- For $i = 0, 1, \ldots$:
  - $Y_{\delta,i} = x + \kappa_{\delta,i} v$ for $\delta = 0, 1$
  - $\delta^* = \arg\min_\delta f(Y_{\delta,i})$
  - $\kappa_{\delta^*,i+1} = \kappa_{\delta^*,i}$
  - $\kappa_{1-\delta^*,i+1} = \mathrm{Mean}(\kappa_{\delta^*,i}, \kappa_{1-\delta^*,i})$

Example: a dichotomic line search



Convergence rates

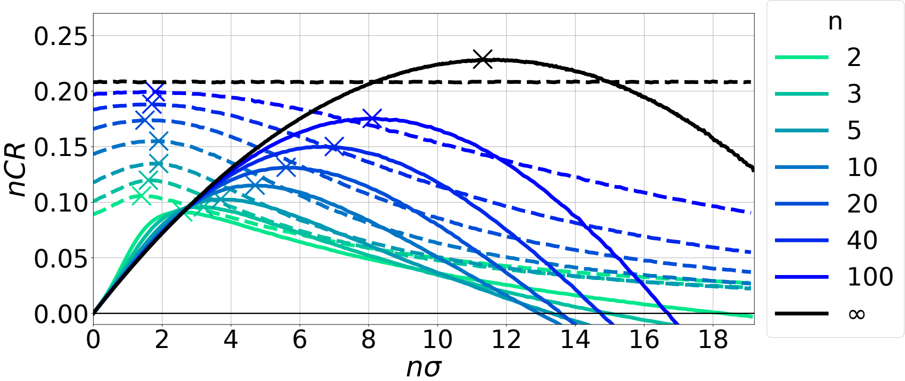


Figure: Convergence rate versus step-size without line search (solid lines) and with dichotomic line search (dashed lines).

Thank you!