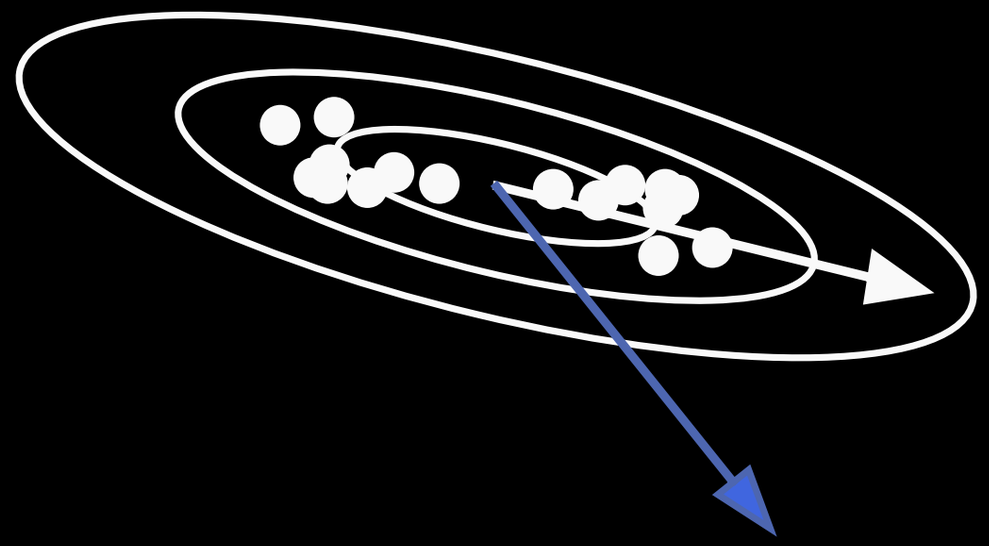


Guided Evolutionary Strategies

Augmenting random search with surrogate gradients

Niru Maheswaranathan // Google Research, Brain Team

Joint work with: Luke Metz, George Tucker, Dami Choi, Jascha Sohl-dickstein

Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Example applications

Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Example applications
- Neural networks with non-differentiable layers

Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Example applications
- Neural networks with non-differentiable layers
- Meta-learning (where computing an exact meta-gradient is costly)

Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Example applications
- Neural networks with non-differentiable layers
- Meta-learning (where computing an exact meta-gradient is costly)
- Gradients from surrogate models (synthetic gradients, black box attacks)

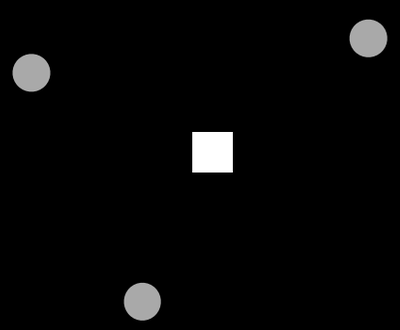
Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

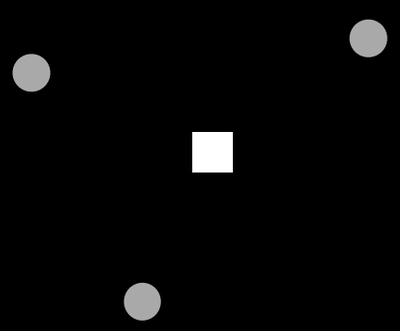
Zeroth-Order
only function values, f(x)



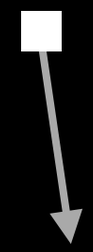
Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

Zeroth-Order
only function values, f(x)



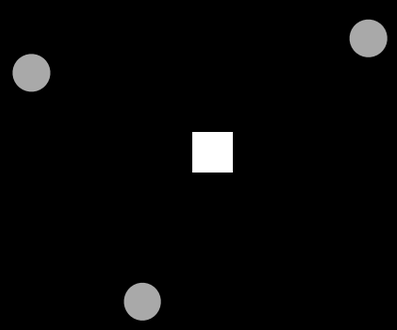
First-Order
gradient information, ∇f(x)



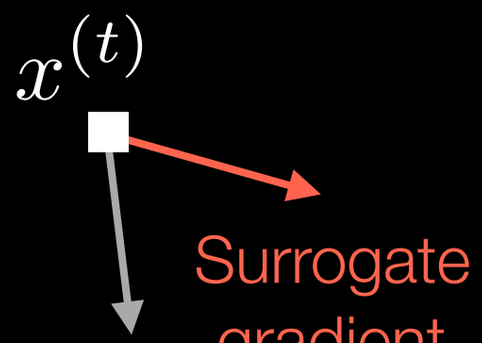
Optimizing with surrogate gradients

Surrogate gradient
directions that are correlated with the true gradient (but may be biased)

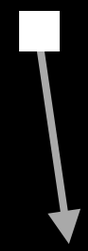
Zeroth-Order
only function values, f(x)



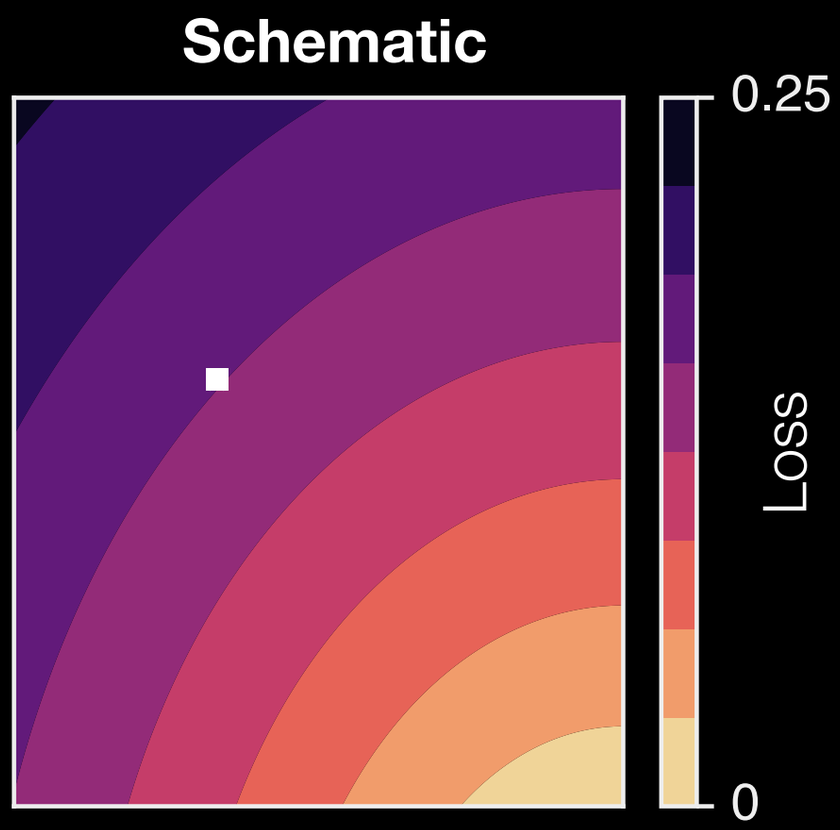
Guided ES



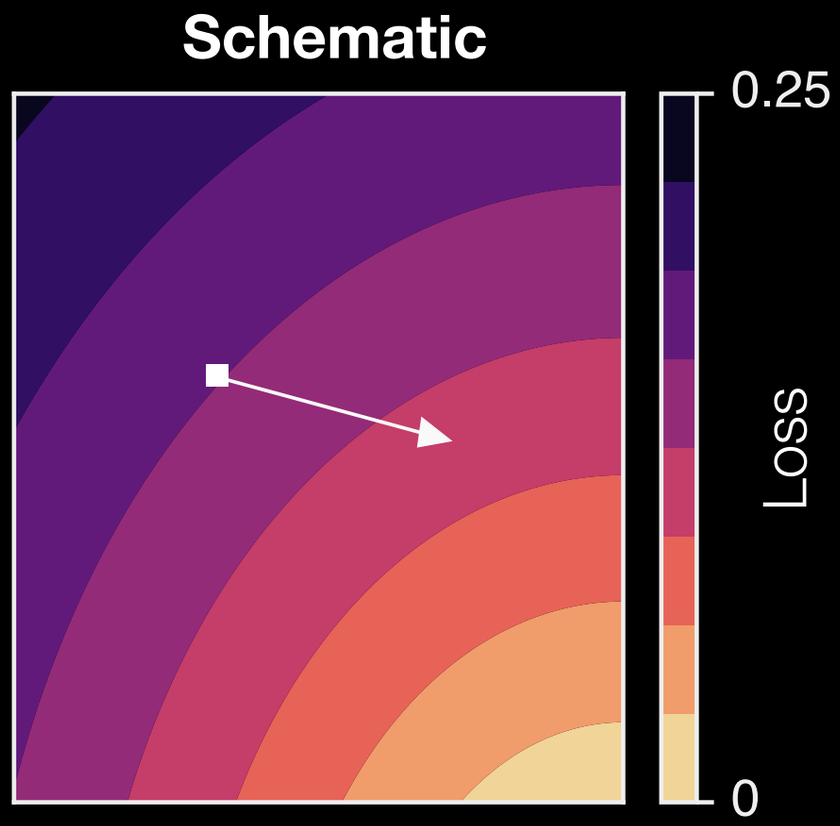
First-Order
gradient information, ∇f(x)



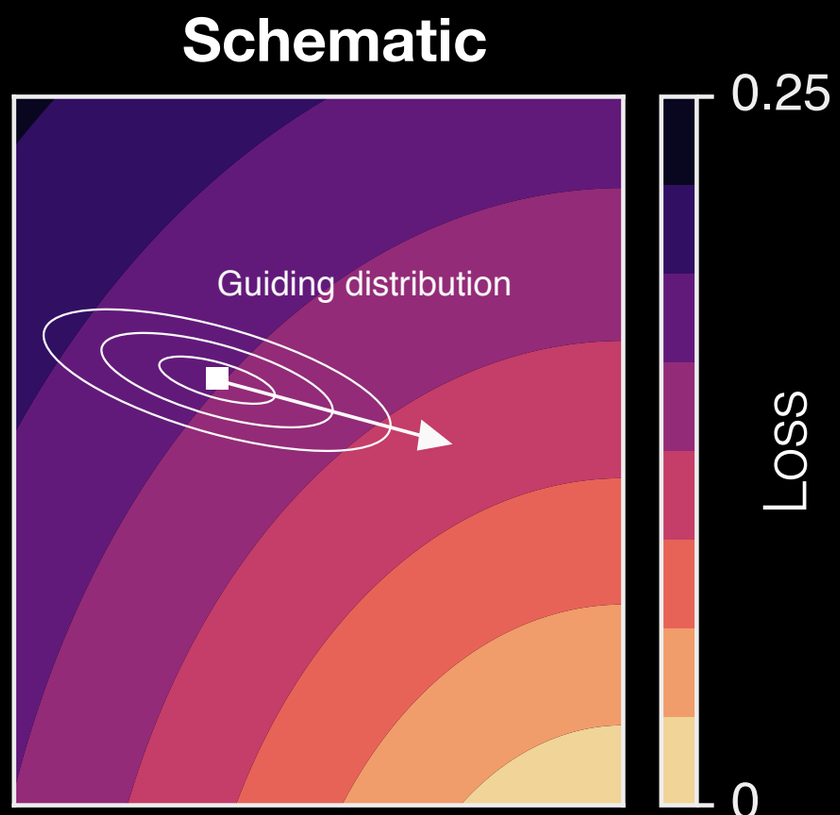
Guided evolutionary strategies



Guided evolutionary strategies

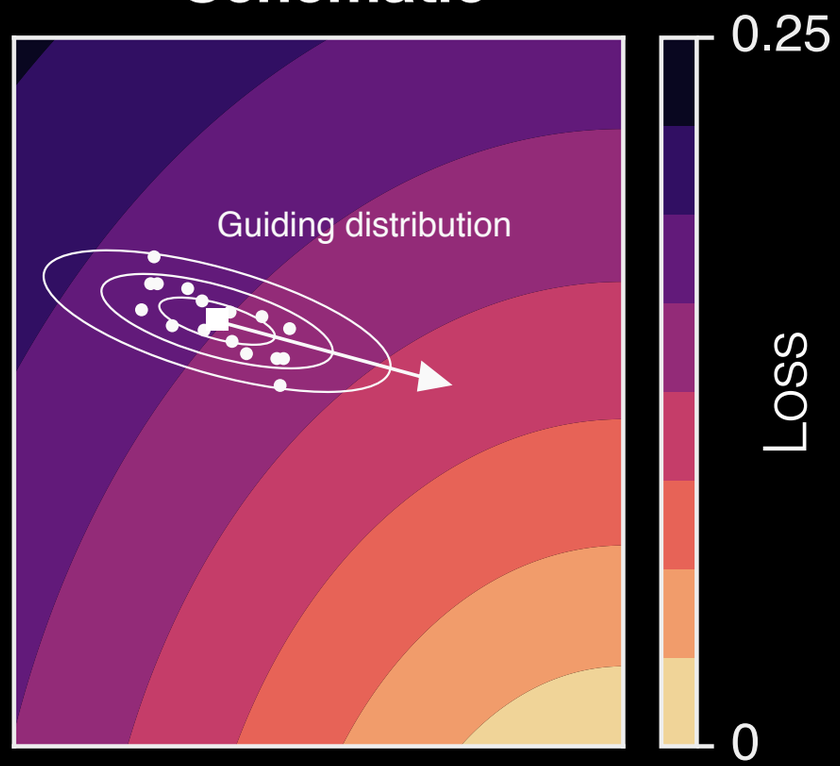


Guided evolutionary strategies



Guided evolutionary strategies

Schematic

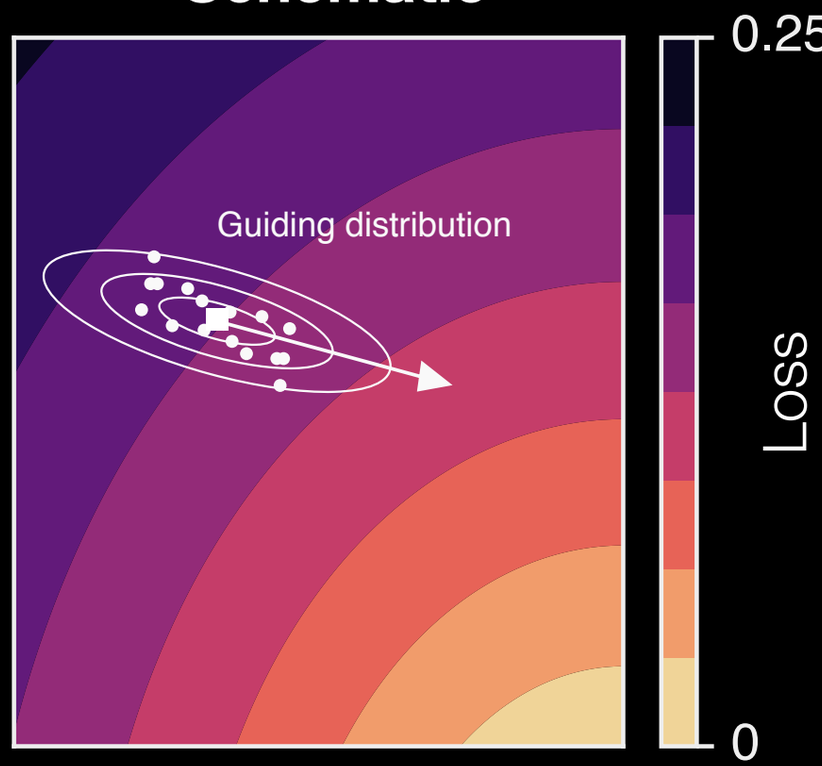


Sample perturbations

$$\epsilon \sim \mathcal{N}(0, \Sigma)$$

Guided evolutionary strategies

Schematic



Sample perturbations

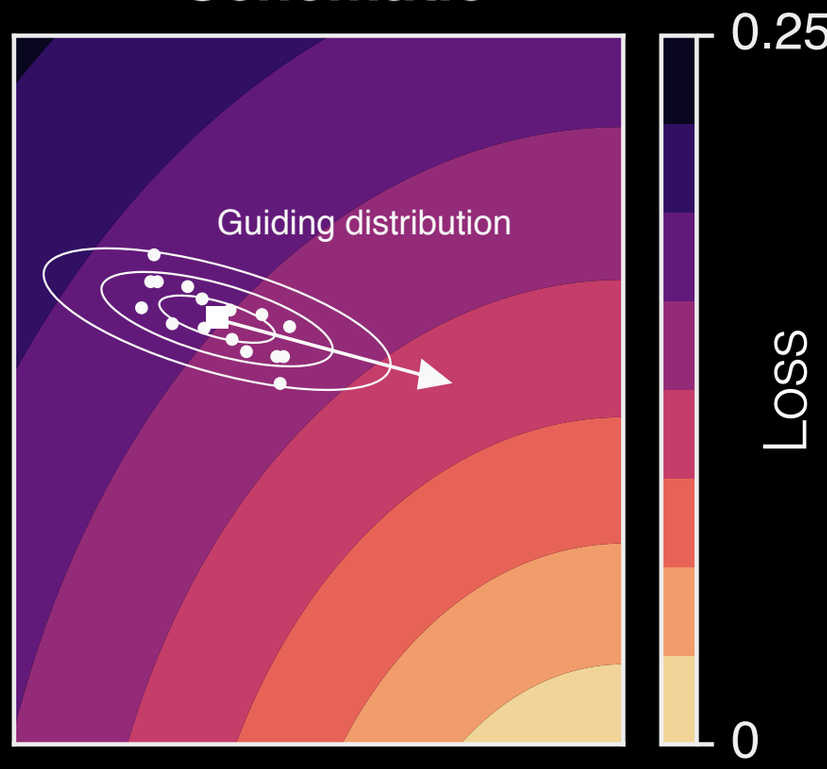
$$\epsilon \sim \mathcal{N}(0, \Sigma)$$

Gradient estimate

$$g = \frac{\beta}{2\sigma^2 P} \sum_{i=1}^{P} \epsilon_i \left(f\left(x + \epsilon_i\right) - f\left(x - \epsilon_i\right) \right)$$

Guided evolutionary strategies

Schematic



Choosing the guiding distribution

Standard (vanilla) ES

Identity covariance

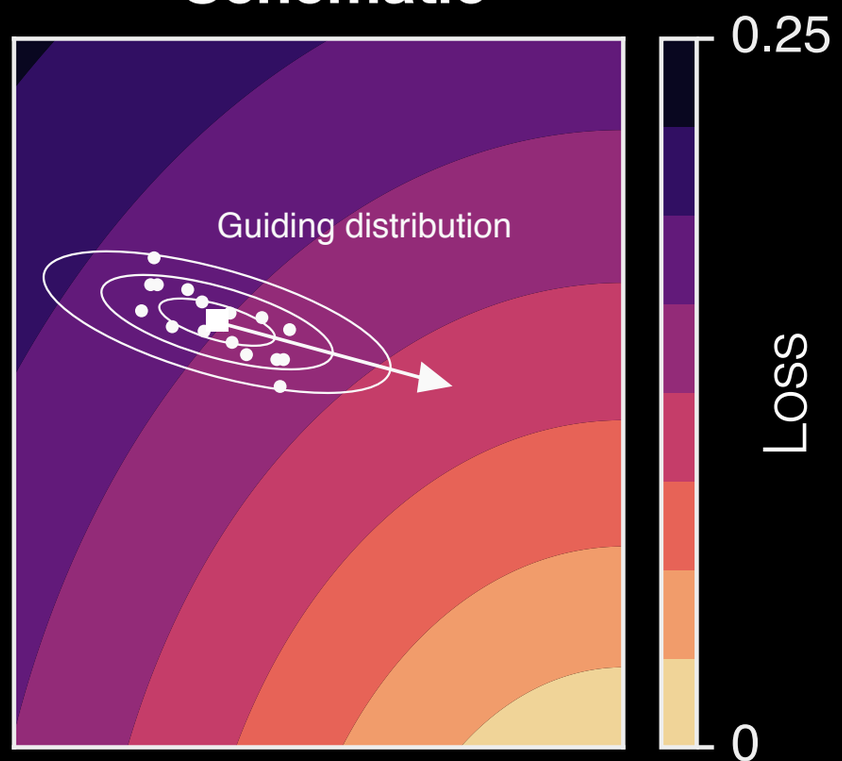
$$\Sigma = \frac{\alpha}{n} I$$

α: hyperparameter

n: parameter dimension

Guided evolutionary strategies

Schematic



Choosing the guiding distribution

Guided ES

Identity + low rank covariance

$$\Sigma = \frac{\alpha}{n} I + \frac{(1-\alpha)}{k} U U^T$$

$$U \in \mathbb{R}^{n \times k}$$

Guiding subspace

columns are surrogate gradients

α: hyperparameter

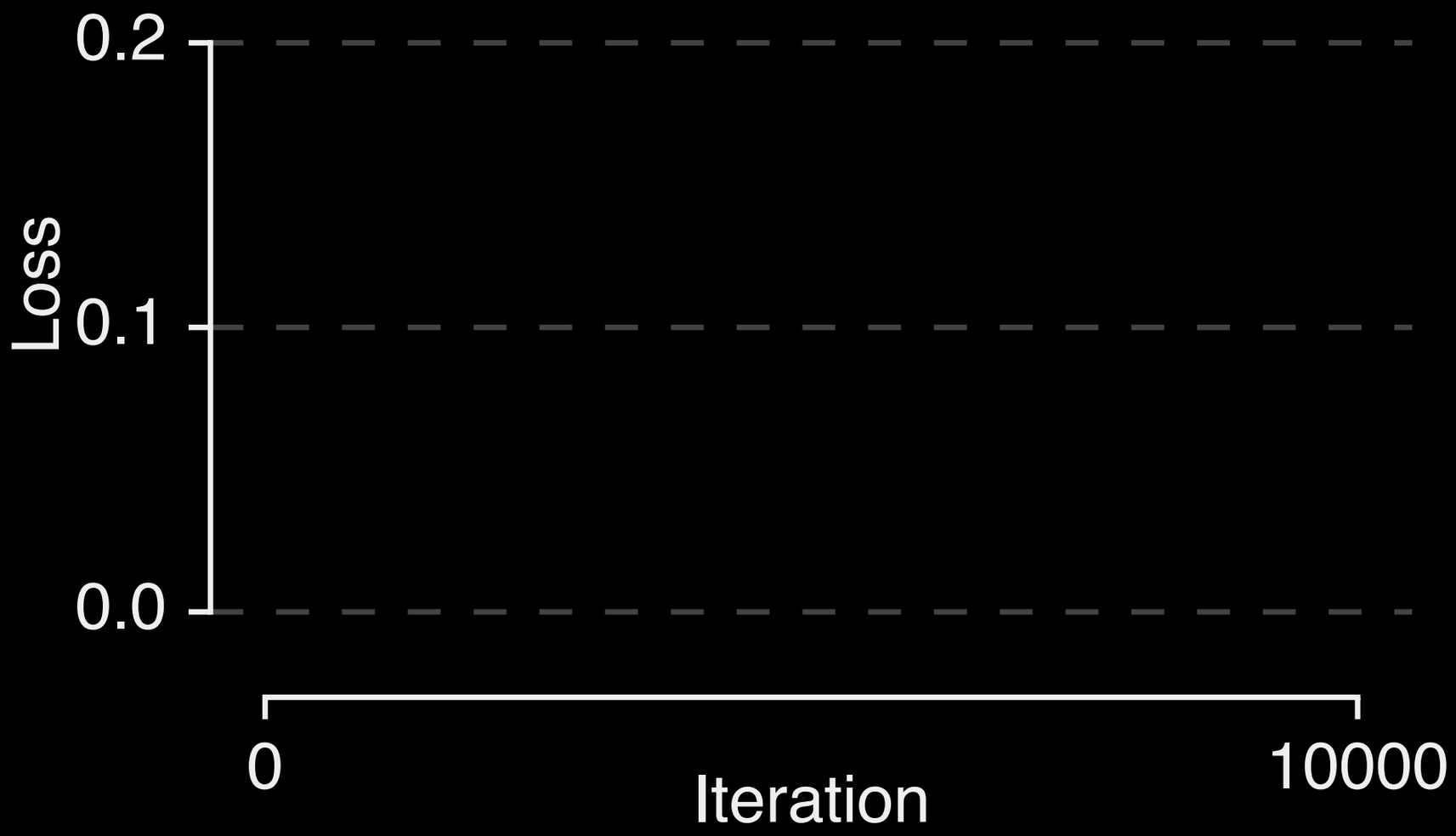
n: parameter dimension

k: subspace dimension

Demo

Perturbed quadratic

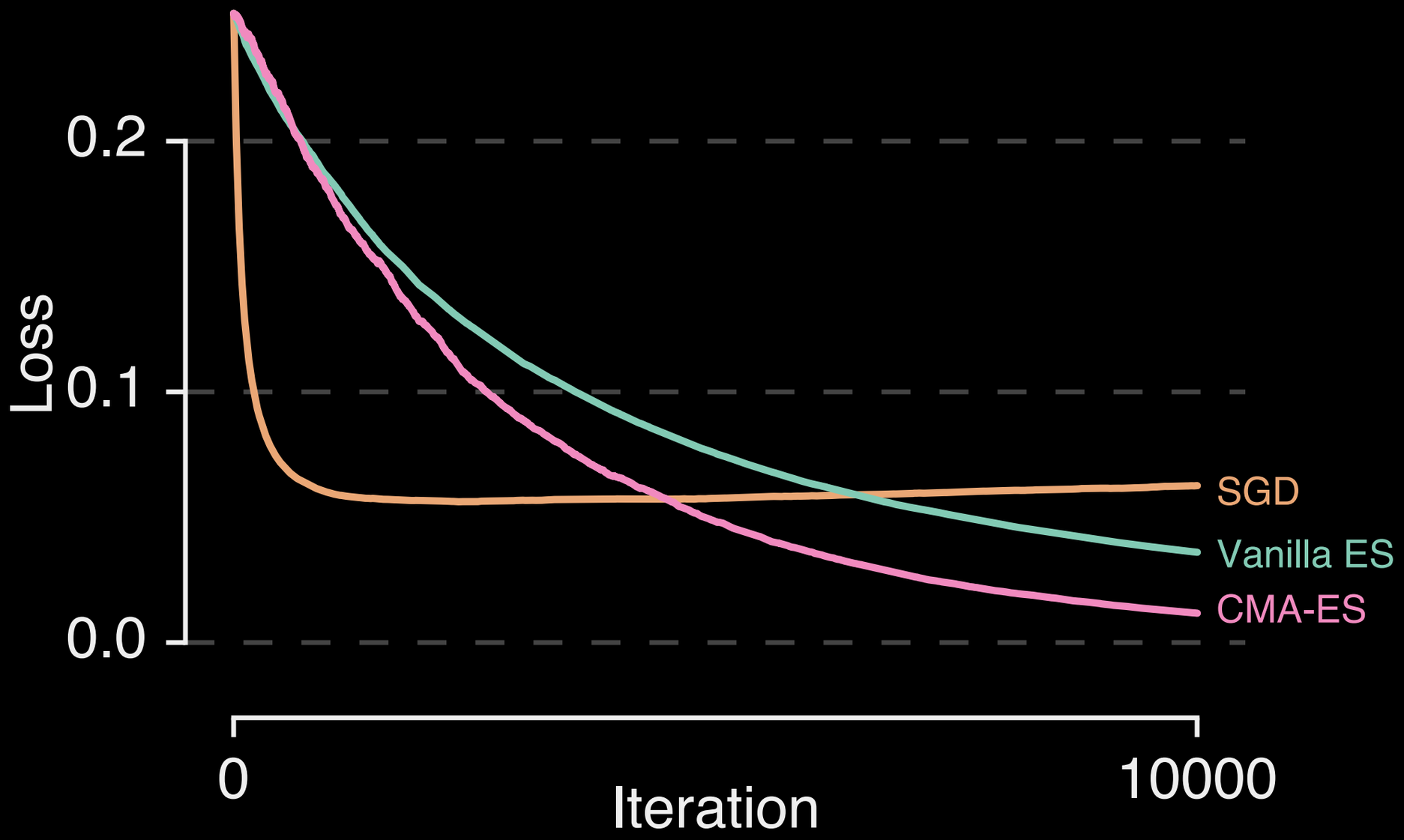
Quadratic function with a bias added to the gradient



Demo

Perturbed quadratic

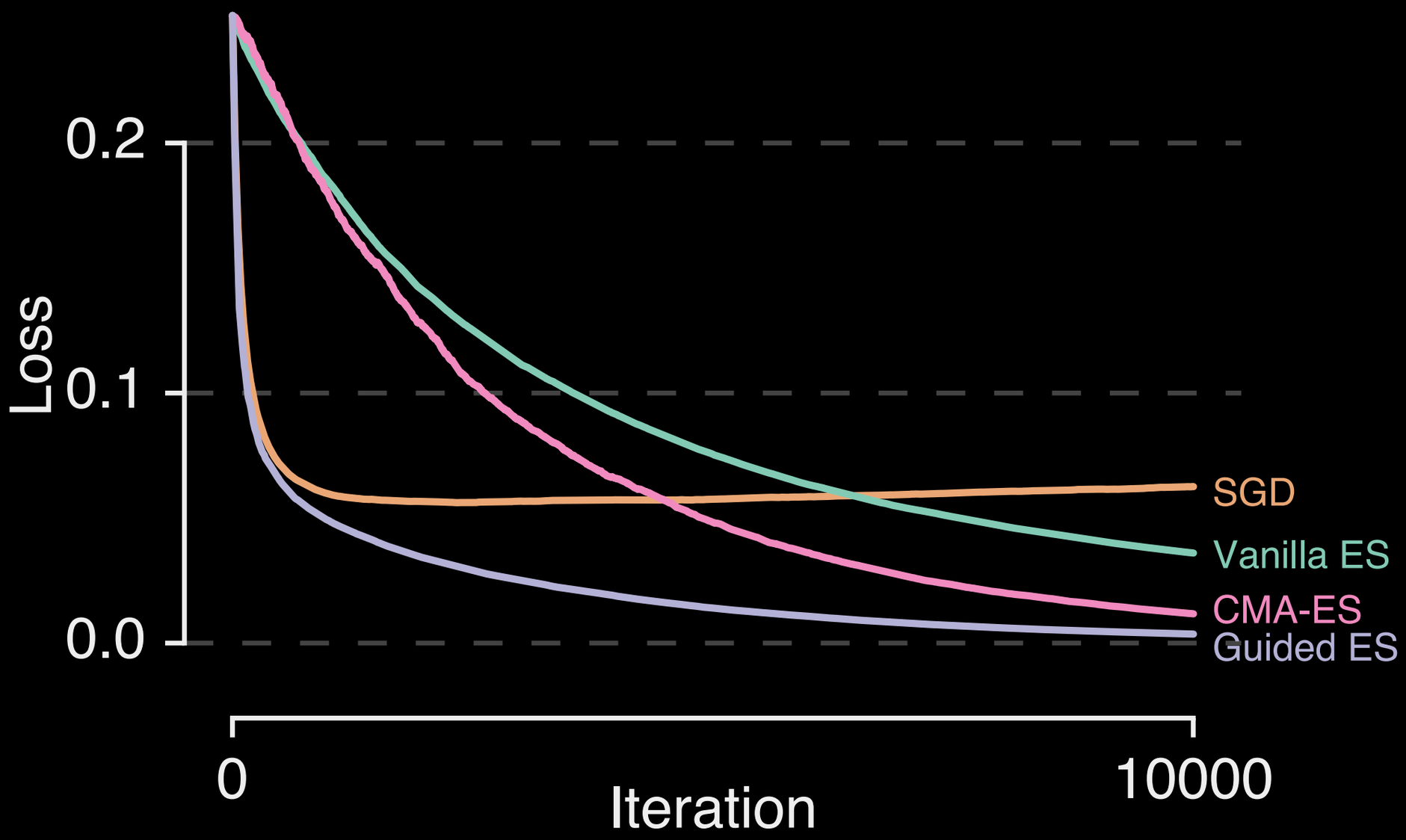
Quadratic function with a bias added to the gradient



Demo

Perturbed quadratic

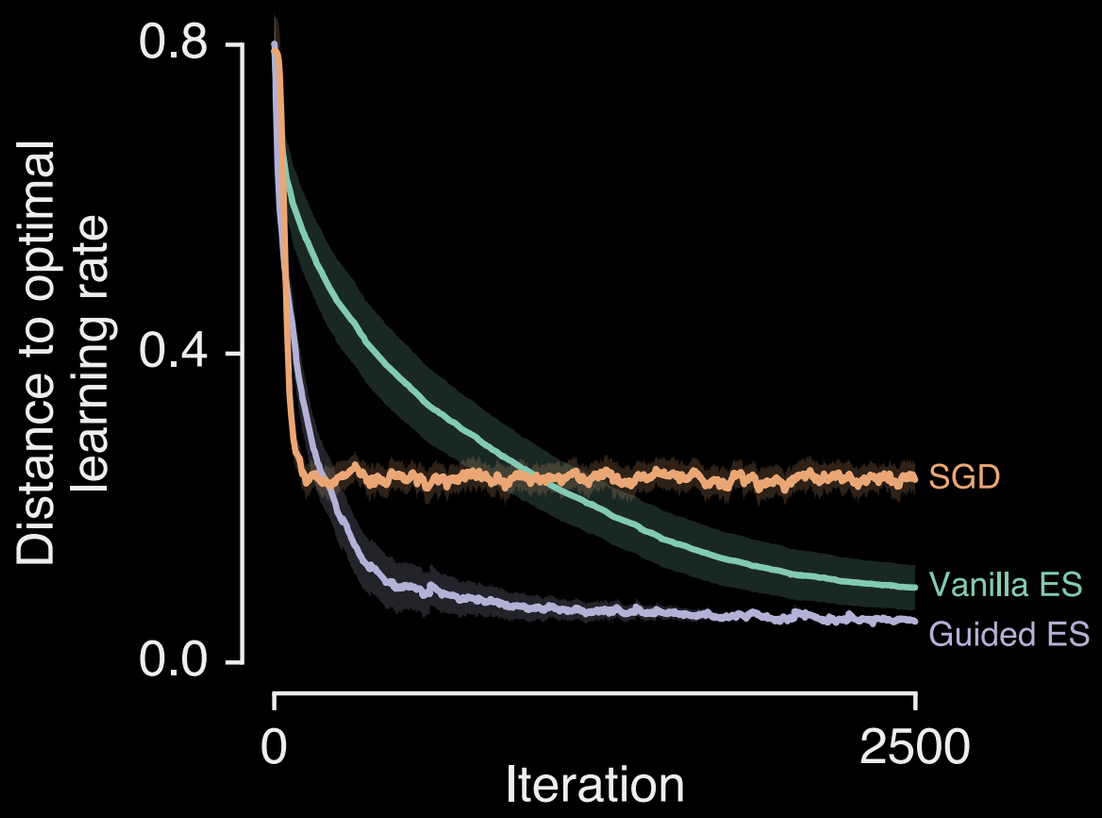
Quadratic function with a bias added to the gradient



Example applications

Unrolled optimization

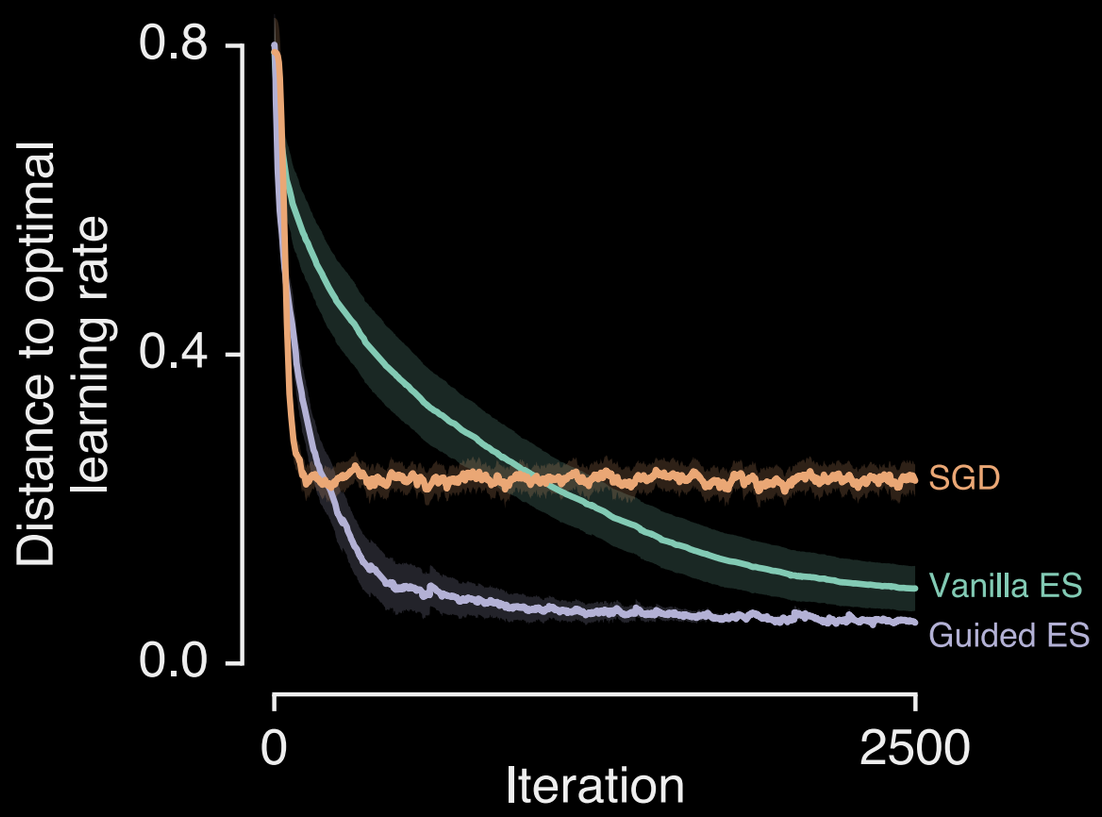
Surrogate gradient from one step of BPTT



Example applications

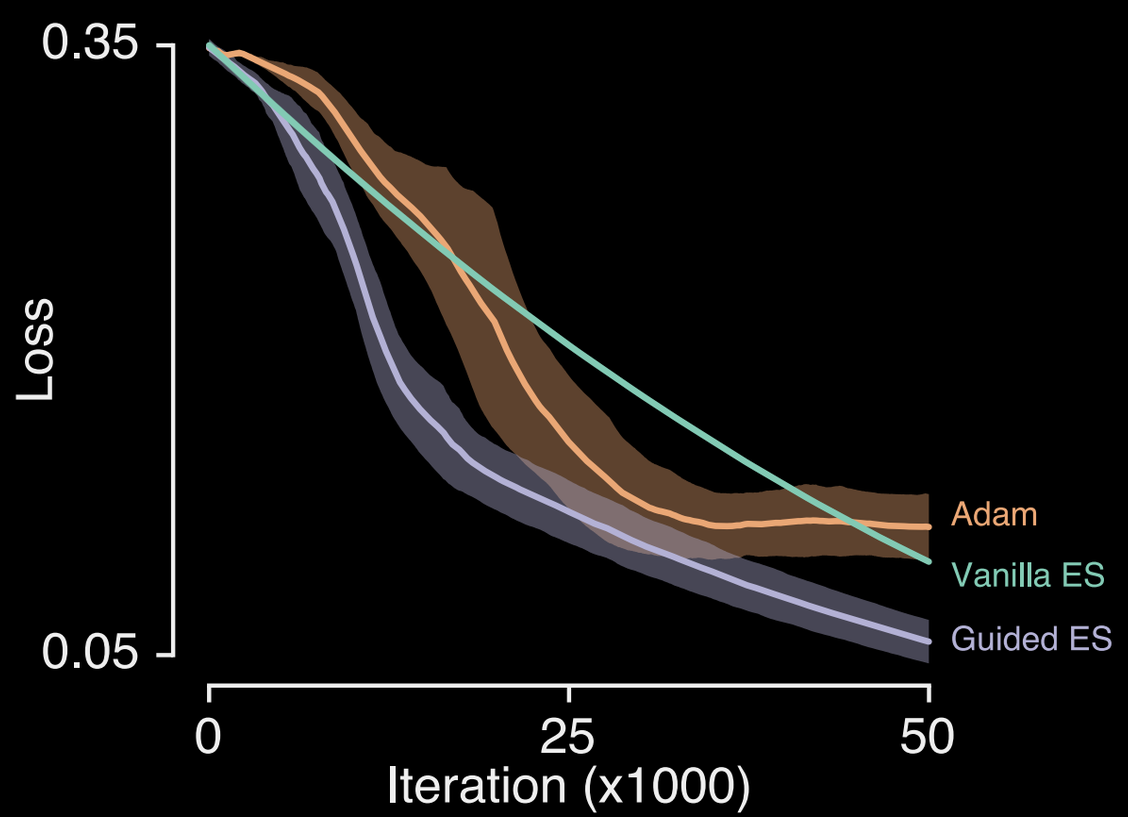
Unrolled optimization

Surrogate gradient from one step of BPTT



Synthetic gradients

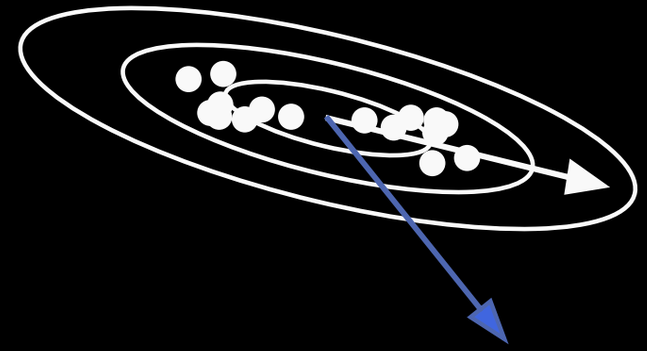
Surrogate gradient is from a synthetic model



Summary

Guided Evolutionary Strategies
Optimization algorithm when you only have access to surrogate gradients

Pacific Ballroom #146
Learn more at our poster



brain-research/guided-evolutionary-strategies

@niru_m

Choosing optimal hyperparameters

Guided ES

Identity + low rank covariance

$$\Sigma = \frac{\alpha}{n} I + \frac{(1-\alpha)}{k} UU^T$$