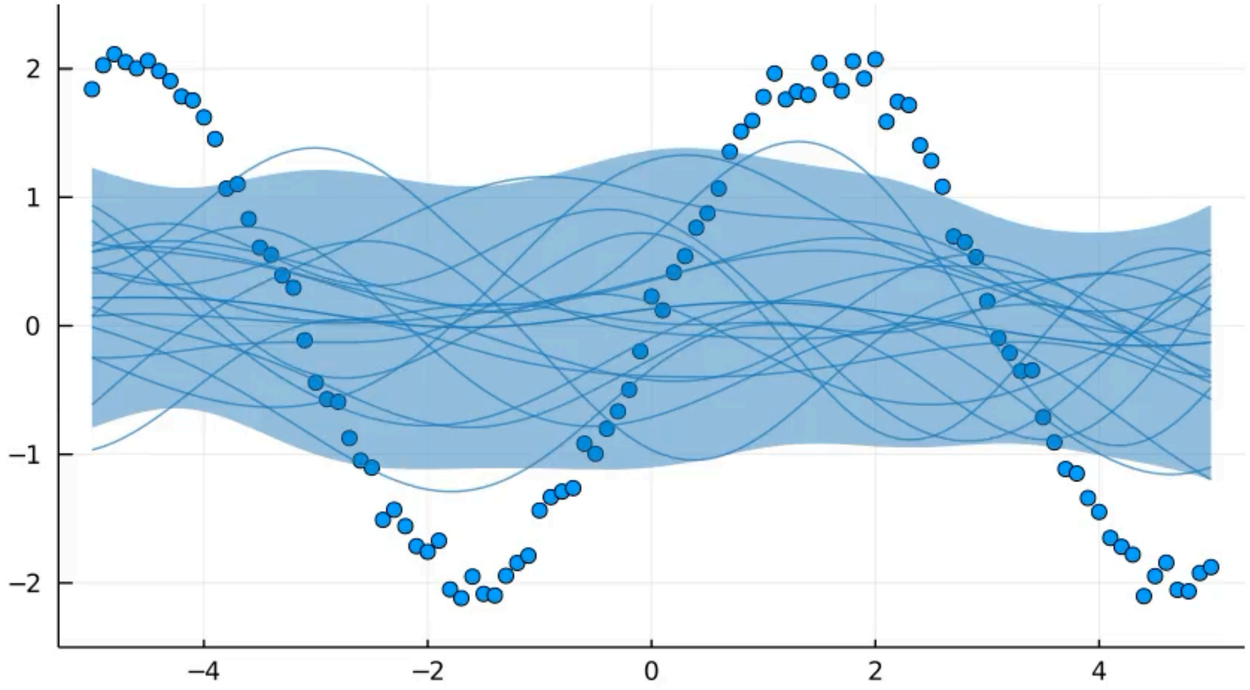
香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen
人工智能學院
School of Artificial Intelligence

Breaking the Curse of Dimensionality in Gaussian Process Training With Zeroth-Order Adaptive Perturbation

Richard Cornelius Suwandi, Feng Yin, Tsung-Hui Chang

Gaussian Process (GP)

GPs are powerful probabilistic models for signal processing and machine learning



The performance of a GP depends *heavily* on its hyperparameters

How to train a GP

Minimize the **negative marginal log-likelihood**:

$$l(\boldsymbol{\theta}) = \frac{1}{2} y^\top C^{-1}(\boldsymbol{\theta}) y + \frac{1}{2} \log\det C(\boldsymbol{\theta})$$

model fit **complexity penalty**

The hyperparameters can be updated as,

$$\boldsymbol{\theta}_i^{t+1} = \boldsymbol{\theta}_i^t - \eta \left.\frac{\partial l(\boldsymbol{\theta})}{\boldsymbol{\theta}_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^t} \quad \text{for all } i = 1,2,\ldots,p$$

gradient

What makes GP training hard

Cubic Complexity

Kernel matrix inversion scales as $O(n^3)$ per iteration, which is prohibitive for large datasets

Numerical Instability

Ill-conditioned kernel matrices yield unreliable gradients and premature convergence

Curse of Dimensionality

High-dimensional landscapes with many local minima

Finite difference stochastic approximation

Approximate the gradients using **finite differences** of function evaluations:

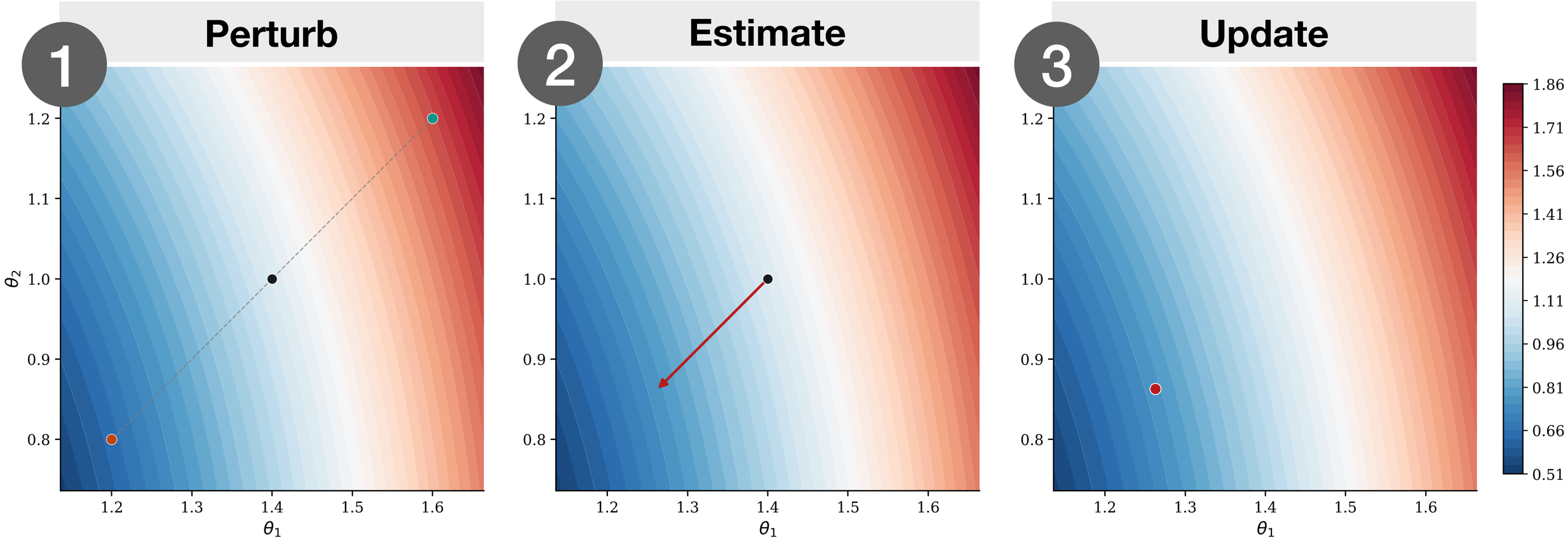
$$g_i(\boldsymbol{\theta}) = \frac{\partial l(\boldsymbol{\theta})}{\boldsymbol{\theta}_i} \approx \frac{l(\boldsymbol{\theta} + h\mathbf{e}_i) - l(\boldsymbol{\theta} - h\mathbf{e}_i)}{2h}$$

where:

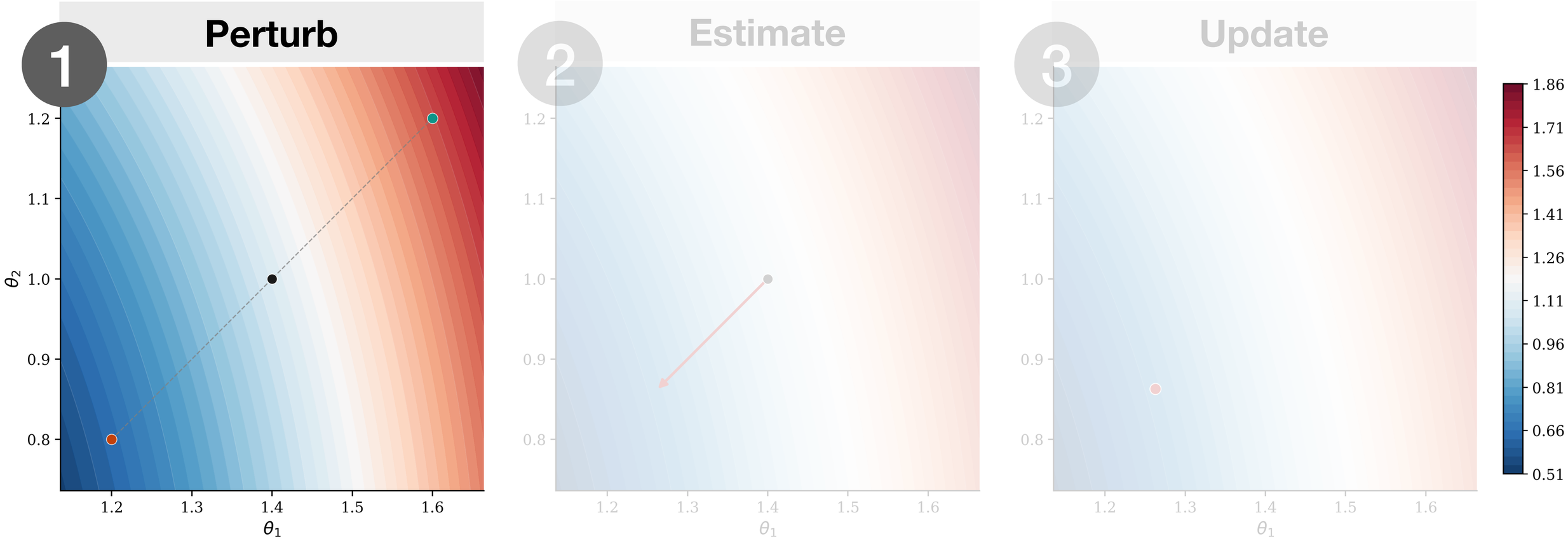
- $\mathbf{e}_i$ is a unit vector in the direction of $\boldsymbol{\theta}_i$
- h is a small step size

Nesterov & Spokoiny, "Random gradient-free minimization of convex functions", Foundations of Computational Mathematics, 2017

Zeroth-order adaptive perturbation (ZAP)

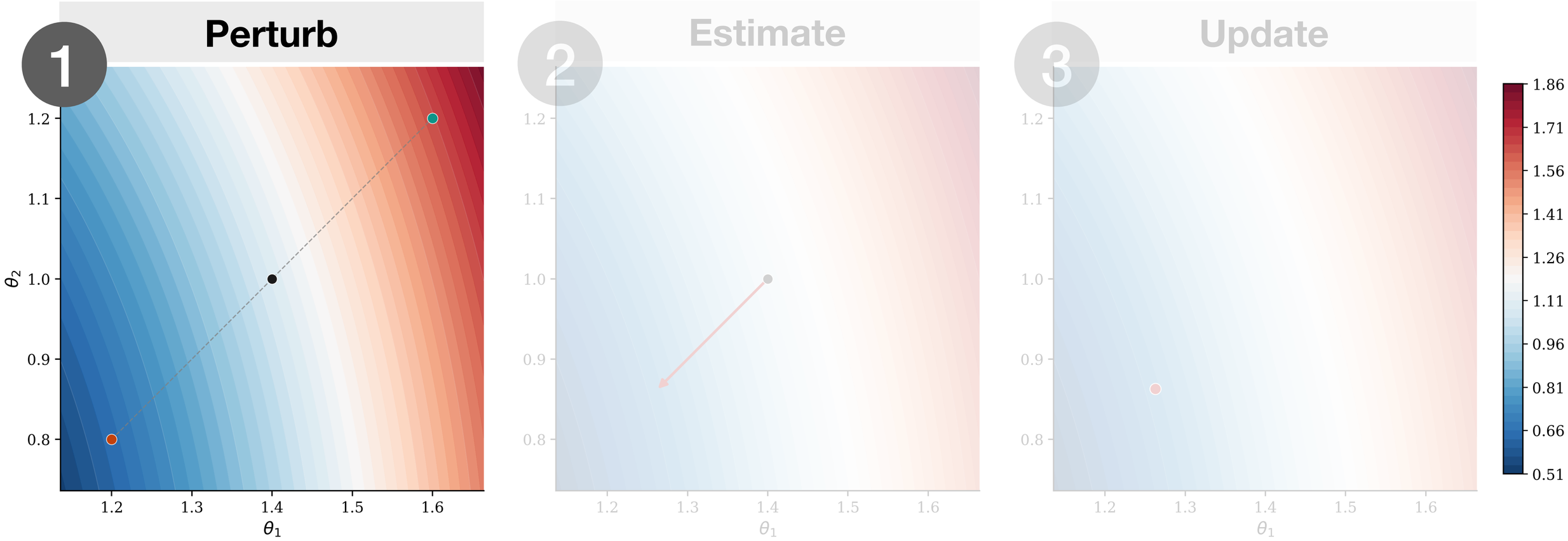


Zeroth-order adaptive perturbation (ZAP)



Draw random vector $\Delta_t \in \{-1,1\}^p$ from symmetric Bernoulli distribution

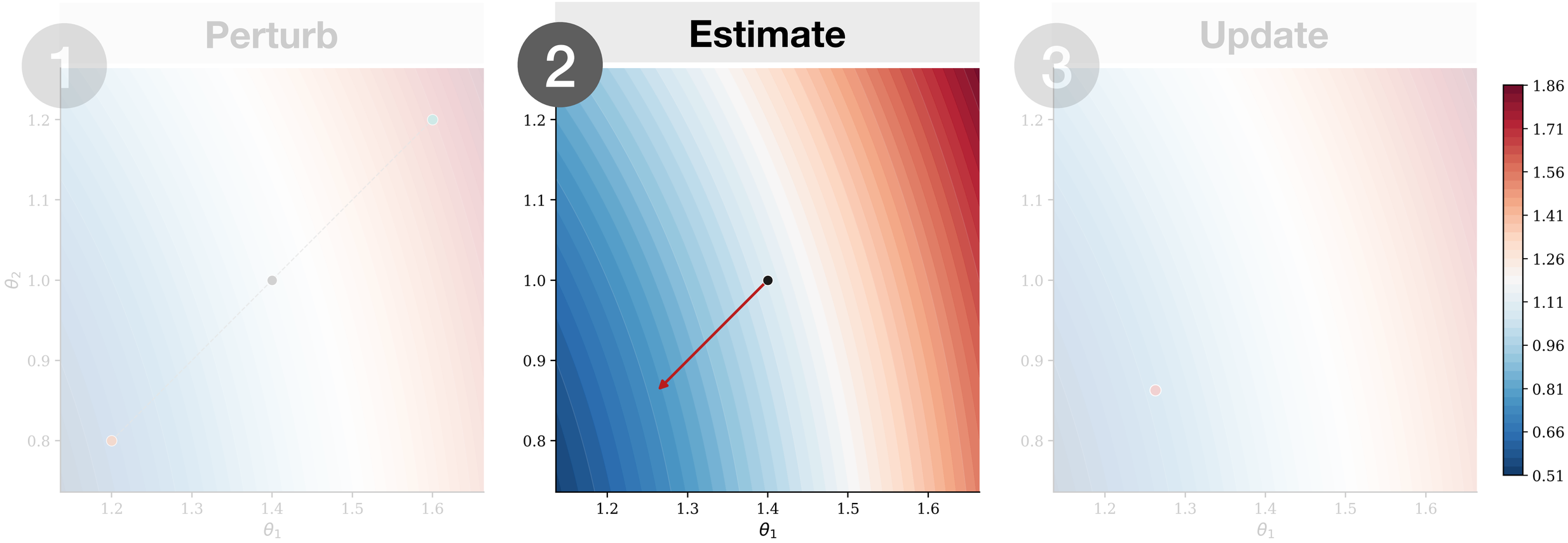
Zeroth-order adaptive perturbation (ZAP)



Evaluate the loss at $\boldsymbol{\theta}_t^+ = \boldsymbol{\theta}_t + \beta_t \Delta_t$ and $\boldsymbol{\theta}_t^- = \boldsymbol{\theta}_t - \beta_t \Delta_t$,

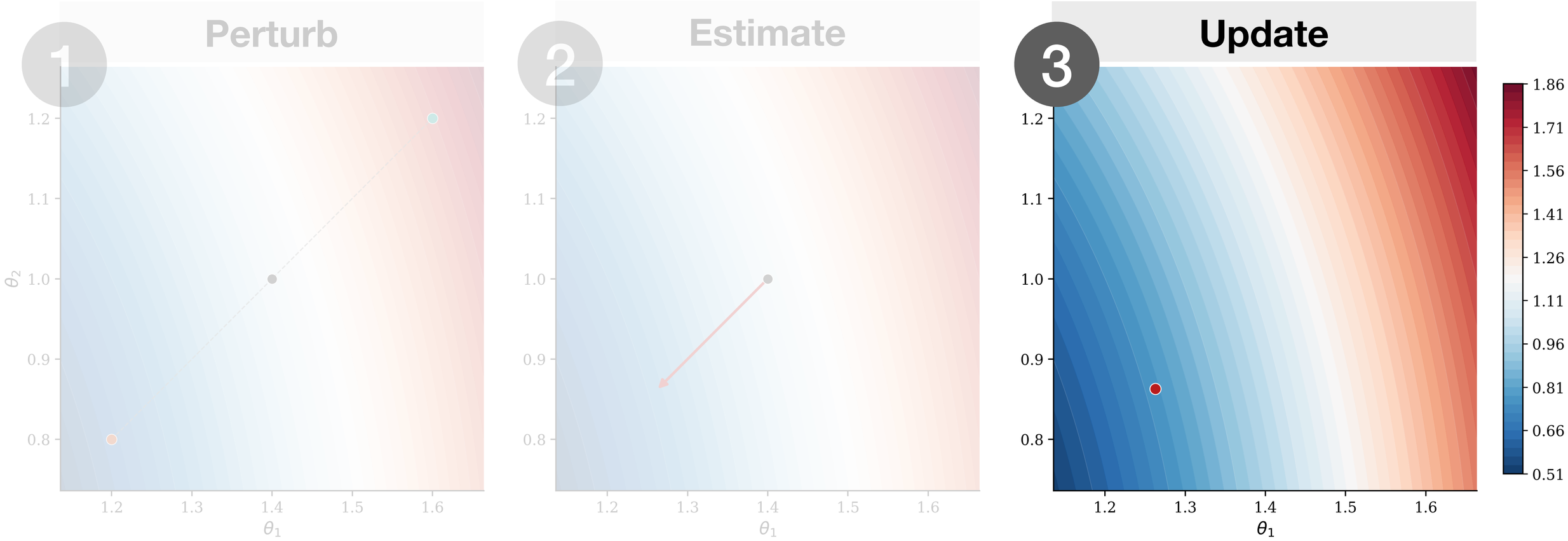
where $\beta_t > 0$ is the perturbation step size

Zeroth-order adaptive perturbation (ZAP)



Estimate the gradient as $\hat{g}_t(\boldsymbol{\theta}_t) = l(\boldsymbol{\theta}_t^+) - l(\boldsymbol{\theta}_t^-)/2\beta_t \cdot \Delta_t^{-1}$

Zeroth-order adaptive perturbation (ZAP)



Update the hyperparameters as $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \alpha_t \hat{g}_t(\boldsymbol{\theta}_t)$, where $\alpha_t > 0$ is the step size

Why ZAP?

Methods	Requires gradient?	Number of evals	Complexity
SGD / Adam	Yes	1	$O(p)$
Finite difference stochastic approximation (FDSA)	No	$2p$	$O(p)$
ZAP	No	2	$O(1)$

Analysis

The step sizes need to satisfy the following assumptions:

Assumption

$$\alpha_t, \beta_t > 0 \text{ for all } t \text{ and } \alpha_t \to 0, \beta_t \to 0 \text{ as } t \to \infty,$$

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \quad \sum_{t=0}^{\infty} \left(\frac{\alpha_t}{\beta_t} \right)^2 < \infty$$

We set a **step size schedule:**

$$\alpha_t = \frac{a}{(A + t + 1)^{\tau}}, \quad \beta_t = \frac{b}{(t + 1)^{\gamma}},$$

where a, A, b, τ, γ are non-negative constants

Analysis

$\hat{g}_t(\cdot)$ is an unbiased estimator of $g_t(\cdot)$ as $t \to \infty$

Lemma (asymptotically unbiased estimator)

For all $\{\boldsymbol{\theta}_t\}$, we have $\mathbb{E}[\hat{g}_t(\boldsymbol{\theta}_t) - g(\boldsymbol{\theta}_t) \,|\, \boldsymbol{\theta}_t] = O(\beta_t^2)$, and $\beta_t \to 0$ as $t \to \infty$

Analysis

$\hat{g}_t(\cdot)$ is an unbiased estimator of $g_t(\cdot)$ as $t \to \infty$

Lemma (asymptotically unbiased estimator)

For all $\{\boldsymbol{\theta}_t\}$, we have $\mathbb{E}[\hat{g}_t(\boldsymbol{\theta}_t) - g(\boldsymbol{\theta}_t) \,|\, \boldsymbol{\theta}_t] = O(\beta_t^2)$, and $\beta_t \to 0$ as $t \to \infty$

ZAP **converges almost surely** to a stationary point

Theorem (almost sure convergence)

Under standard SA assumptions, $\boldsymbol{\theta}_t \to \boldsymbol{\theta}^*$ almost surely as $t \to \infty$

Experiment setup - model

Automatic Relevance Determination (ARD) kernel

$$k(\mathbf{x}, \mathbf{x}'; \boldsymbol{\theta}_h) = \sigma_f^2 \exp\left(-\frac{1}{2}\sum_{i=1}^{d}\frac{(x_i - x_i')^2}{\ell_i^2}\right)$$

where:

- σ_f^2 is the signal variance
- ℓ_i is the lengthscale for the i-th input dimension
- $\boldsymbol{\theta}_h = [\sigma_f^2, l_1, \ldots, l_d]^\top$, so number of hyperparameters is $p = d + 1$

Experiment setup - baselines

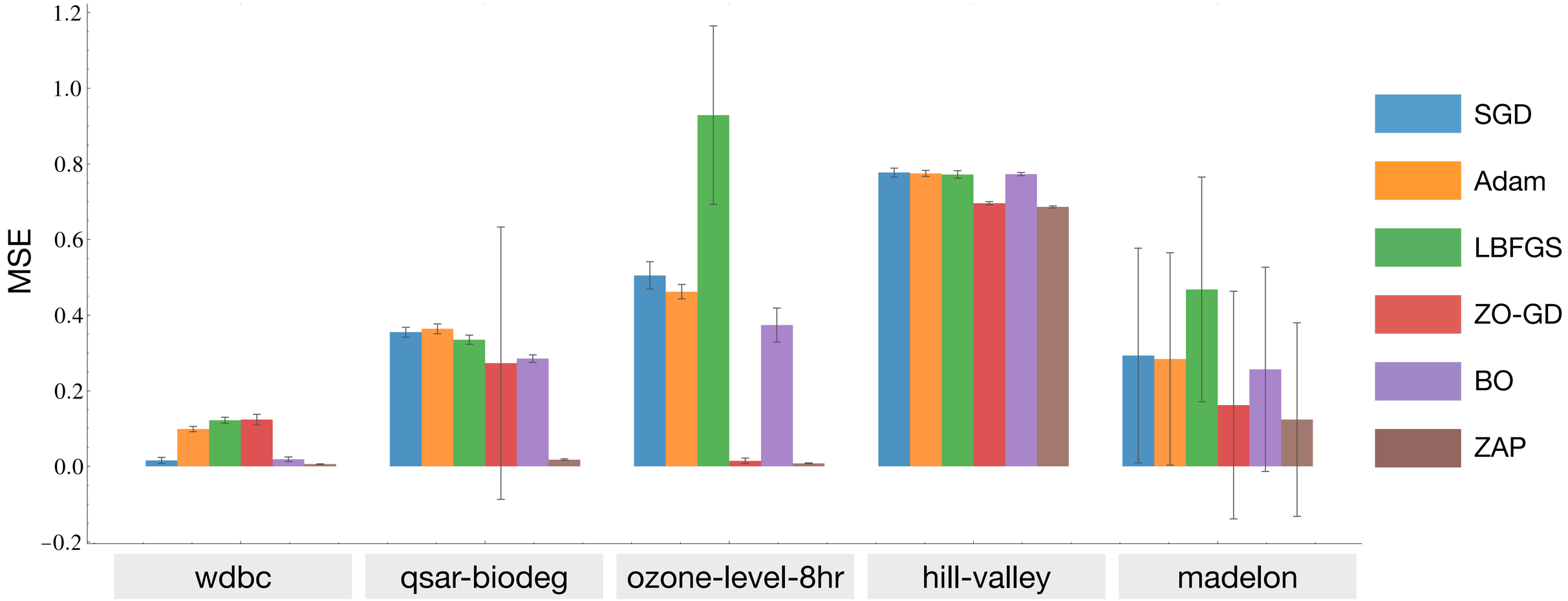
Evaluate ZAP against the following methods:

Gradient-based	**Gradient-free**
Stochastic gradient descent (SGD)	Finite difference stochastic approximation (FDSA)
Adam	Bayesian optimization (BO)
L-BFGS	

Experiment setup - dataset

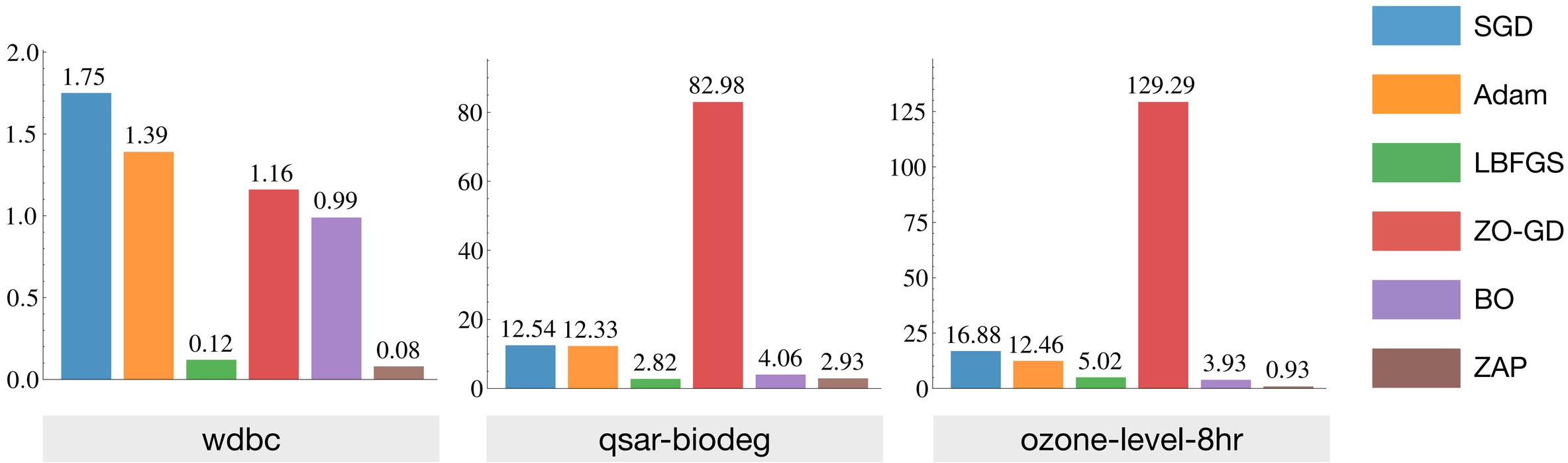
Dataset	Dimensionality (d)
wdbc	30
qsar-biodeg	31
ozone-level-8hr	72
hill-valley	100
madelon	500

Results - optimization performance



ZAP achieves the **lowest** MSE and **scales** effectively to high dimensions

Results - optimization time



ZAP achieves the **fastest** optimization time

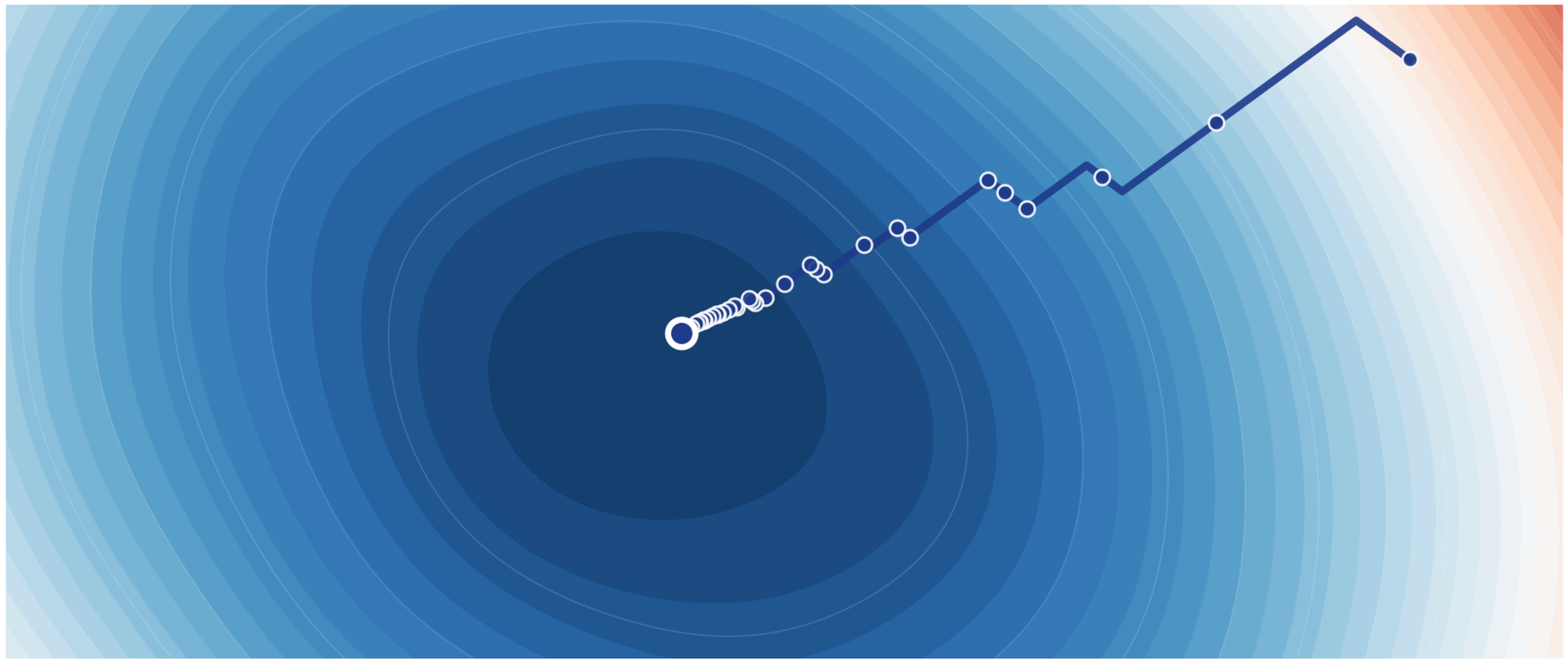
Future work

- Current work only consider the NLML loss
- Extend ZAP to other criteria where the gradients may be unavailable or intractable

Cross-validation	Variational lower bound	Pseudo-likelihood
Leave-one-out cross validation	Evidence lower bound (ELBO)	Laplace approximation
K-fold cross validation	Variational free energy (VFE)	Expectation propagation (EP)

Future work

- Estimated gradients could be noisy due to randomness in perturbations
- Implement a gradient smoothing technique by averaging the gradients across iterations



Thank you for listening!

https://ieeexplore.ieee.org/document/11463847

https://github.com/richardcsuwandi/zap

richardsuwandi@link.cuhk.edu.cn