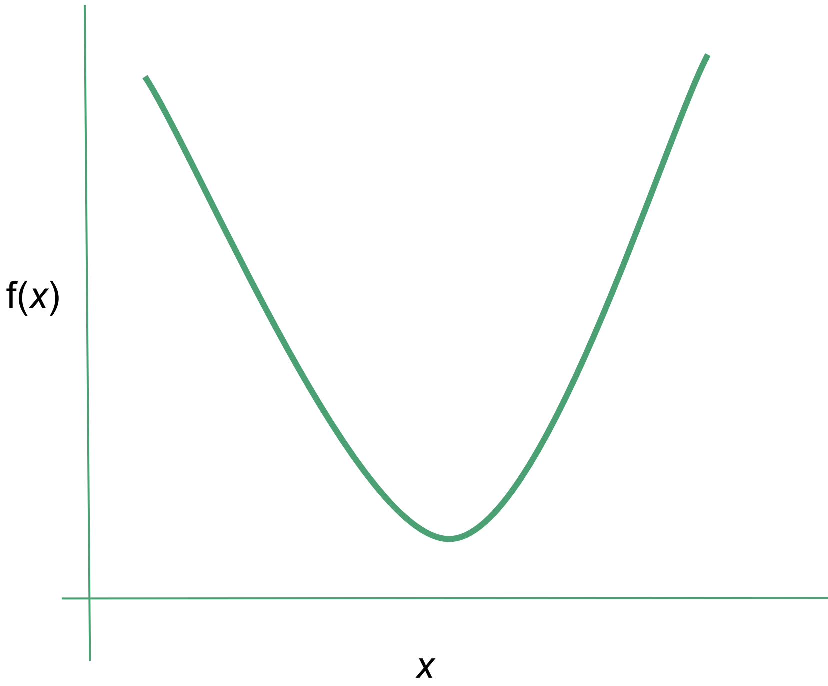
Non-convex optimization

Issam Laradji

Strongly Convex

Objective function

$$\min_{x \in \mathbb{R}^n} \; f(x),$$



Strongly Convex

Objective function

$$\min_{x \in \mathbb{R}^n} \quad f(x),$$

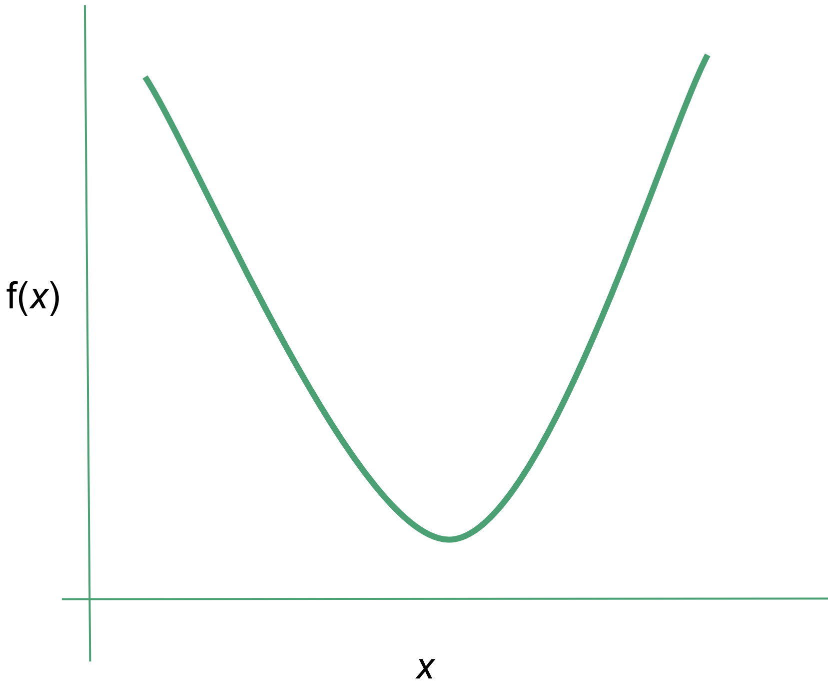
Assumptions

Gradient Lipschitz continuous

$$f(y) \leq f(x) + \nabla f(x)(y - x) + \frac{L}{2}(y - x)^2$$

Strongly convex

$$f(y) \geq f(x) + \nabla f(x)(y - x) + \frac{\mu}{2}(y - x)^2$$



Strongly Convex

Objective function

$$\min_{x \in \mathbb{R}^n} \ f(x),$$

Assumptions

Gradient Lipschitz continuous

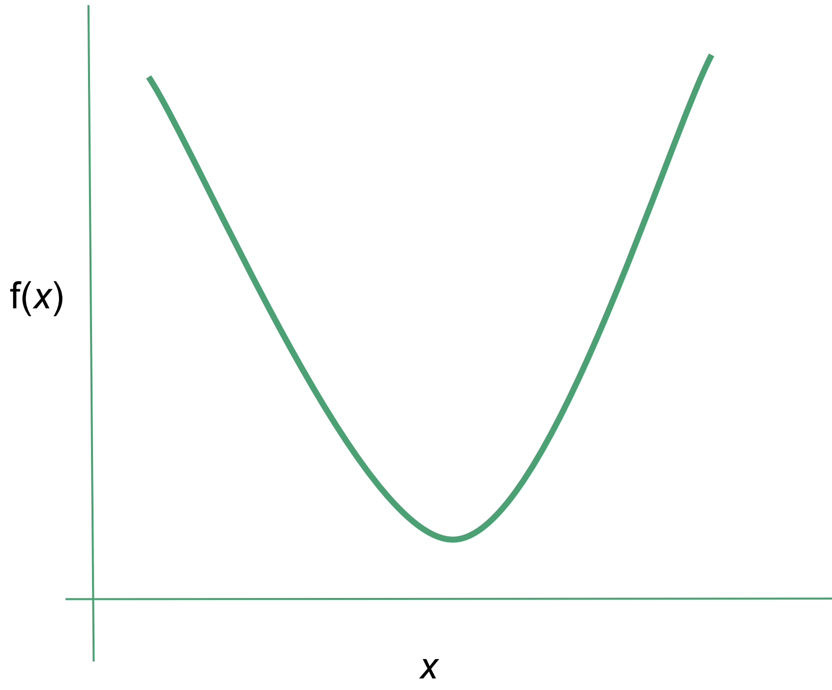
$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$

Strongly convex

$$f(y) \geq f(x) + \nabla f(x)(y-x) + \frac{\mu}{2}(y-x)^2$$

Randomized coordinate descent

$$E[f(x^{t+1}) - f(x^t)] \leq (1 - \frac{\mu}{nL})[f(x^t) - f(x^*)]$$

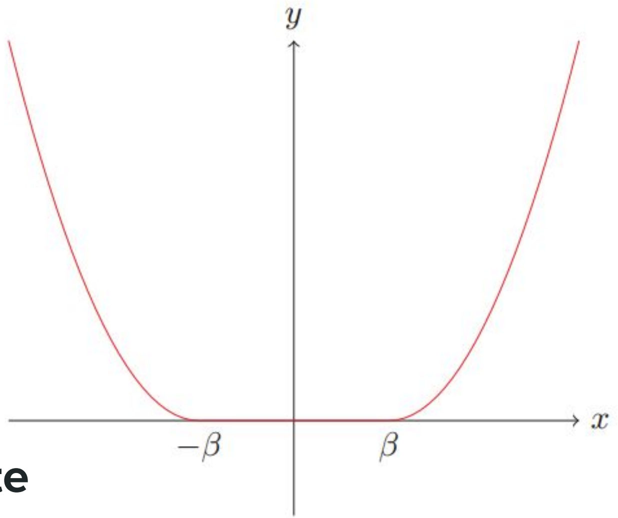


Non-strongly Convex optimization

Assumptions

Gradient Lipschitz continuous

$$f(y) \le f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$



Convergence rate

$$f(x^t) - f(x^*) = O(1/t).$$

Compared to the strongly convex convergence rate

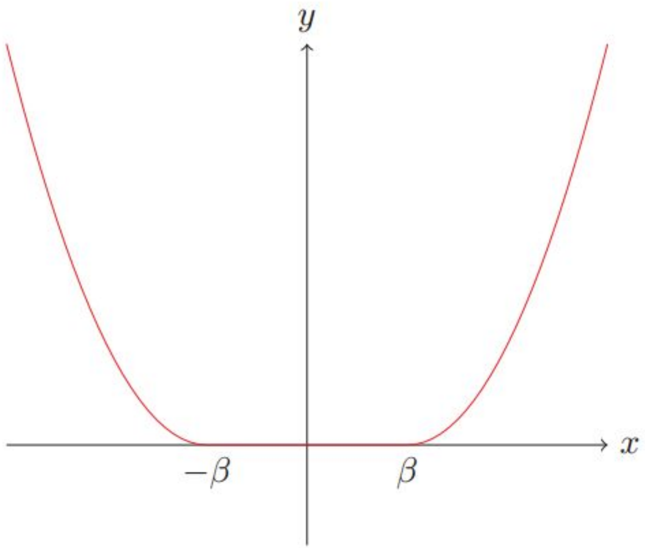
$$f(x^t) - f(x^*) = O(\rho^t)$$

$$E[f(x^{t+1}) - f(x^t)] \le \boxed{(1 - \frac{\mu}{nL})}[f(x^t) - f(x^*)]$$

Non-strongly Convex optimization

Definition 2 (Restricted secant inequality – RSI(ν)). *A function $f(x) : \mathbb{R}^n \to \mathbb{R}$ satisfies the restricted secant inequality (RSI) with constant $\nu > 0$ if it is differentiable and obeys*

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}} \rangle \geq \nu \|x - x_{\text{prj}}\|^2, \quad (7)$$



Non-Strongly Convex

Objective function

$$\min_{x \in \mathbb{R}^n} \quad f(x),$$

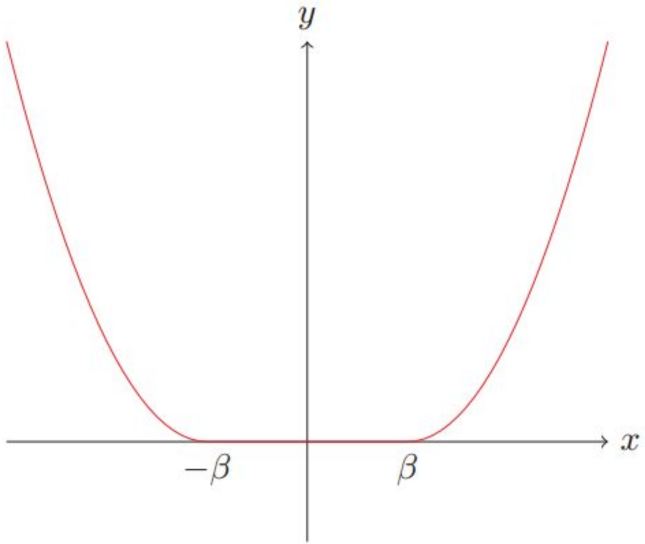
Assumptions

Lipschitz continuous

$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$

Restricted secant inequality

$$\langle \nabla f(x) - \nabla f(x_{\text{prj}}), x - x_{\text{prj}} \rangle \geq \nu \|x - x_{\text{prj}}\|^2.$$



Randomized coordinate descent

$$E[f(x^{t+1}) - f(x^t)] \leq (1 - \frac{\nu}{nL})[f(x^t) - f(x^*)]$$

Invex functions (a generalization of convex function)

Objective function

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

Assumptions

Lipschitz continuous

$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$

Polyak [1963]

$$||\nabla f(x)||^2 \geq 2\mu(f(x) - f^*),$$

This inequality simply requires that the gradient grows faster than a linear function as we move away from the optimal function value.

Invex function (one global minimum)

Invex functions (a generalization of convex function)

Objective function

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

Assumptions

Lipschitz continuous

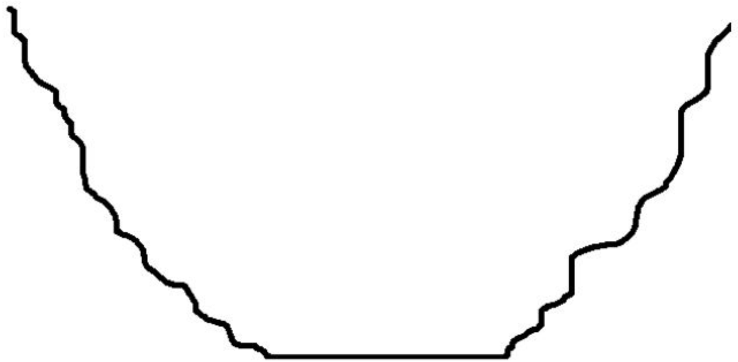
$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$

Polyak [1963] - for invex functions where this holds

$$||\nabla f(x)||^2 \geq 2\mu(f(x) - f^*),$$

Randomized coordinate descent

$$E[f(x^{t+1}) - f(x^t)] \leq (1 - \frac{\mu}{nL})[f(x^t) - f(x^*)]$$



Invex function (one global minimum)

Invex functions (a generalization of convex function)

Objective function

$$\min_{x \in \mathbb{R}^n} \; f(x),$$

Assumptions

Lipschitz continuous

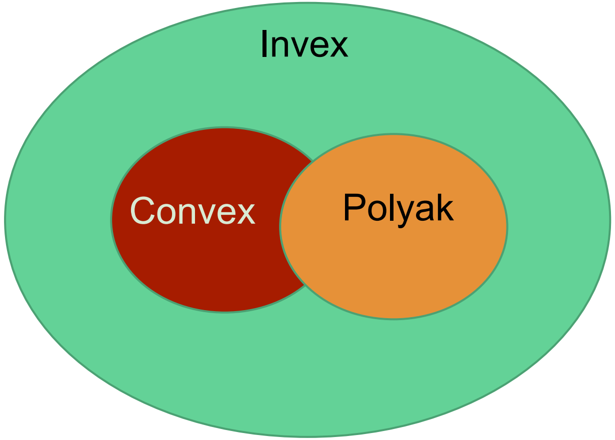
$$f(y) \le f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$

Polyak [1963] - for invex functions where this holds

$$||\nabla f(x)||^2 \ge 2\mu(f(x) - f^*),$$

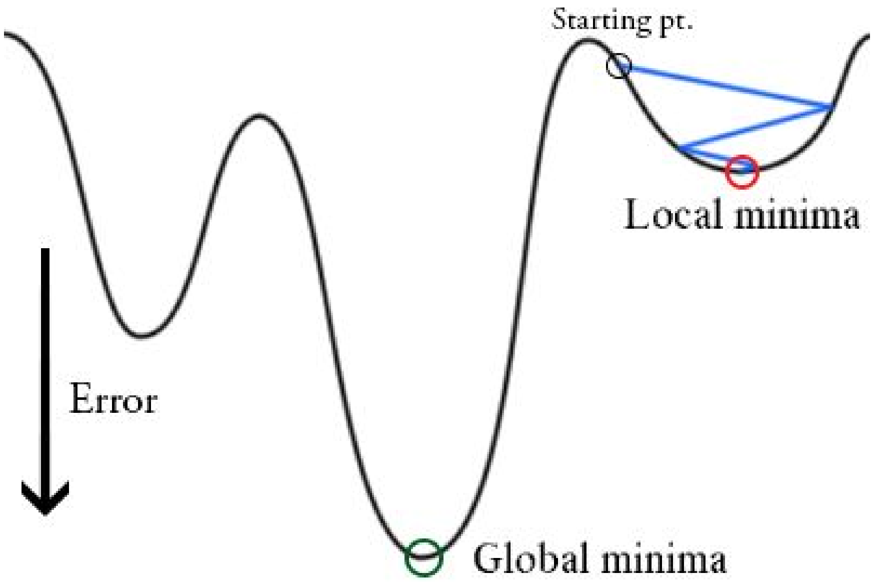
Randomized coordinate descent

$$E[f(x^{t+1}) - f(x^t)] \le (1 - \frac{\mu}{nL})[f(x^t) - f(x^*)]$$

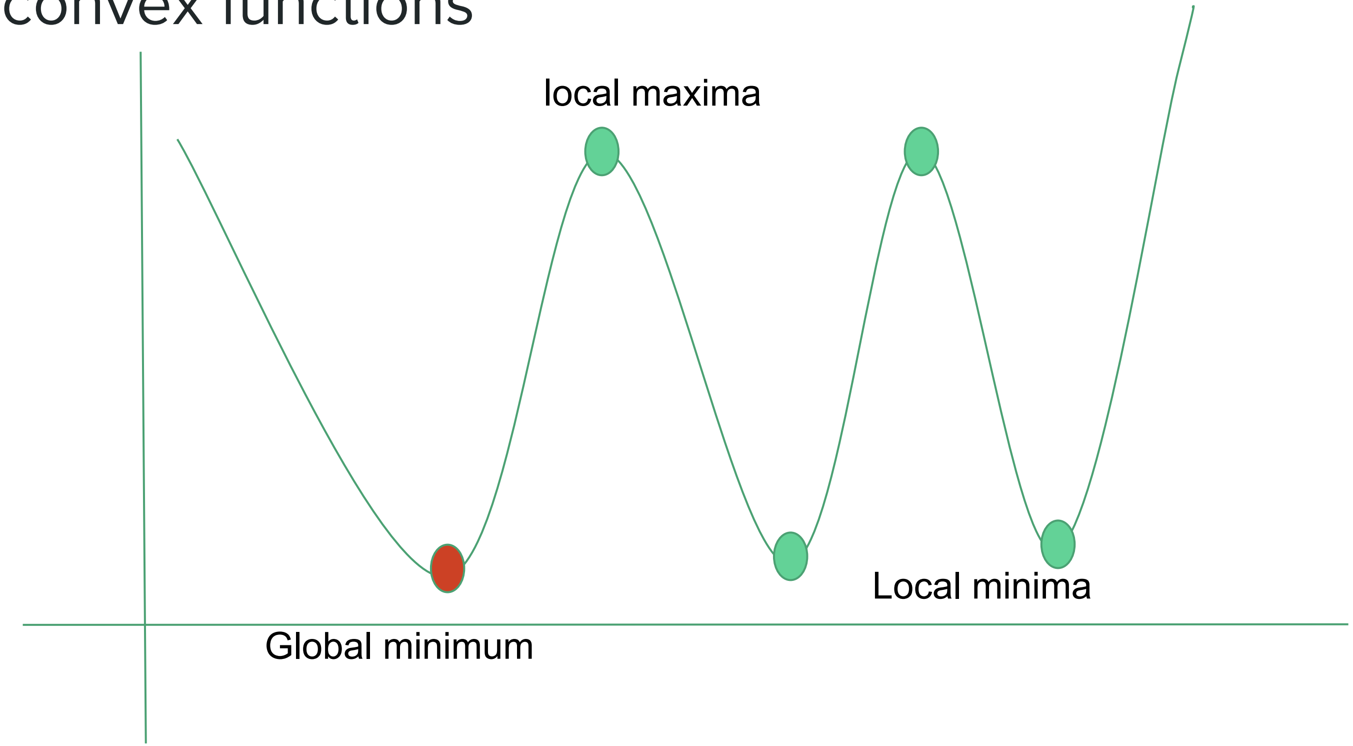


Venn diagram

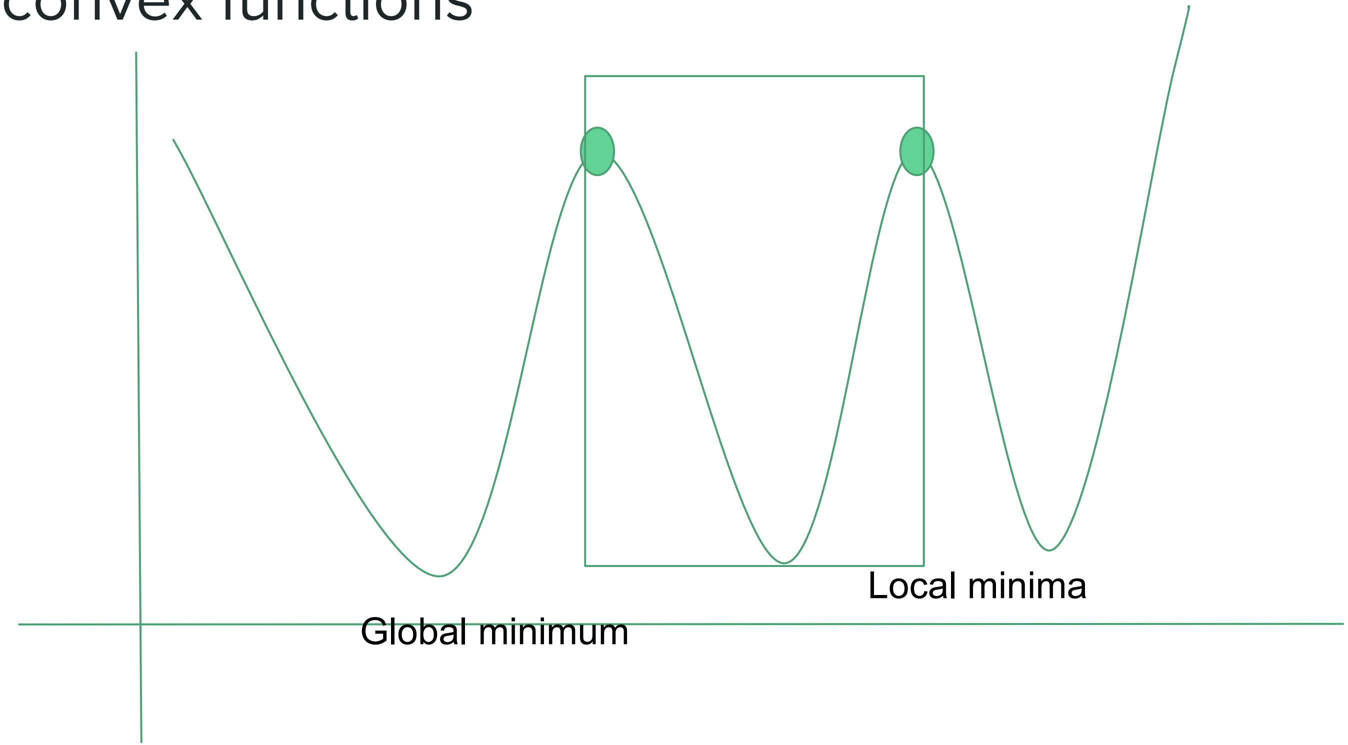
Non-convex functions



Non-convex functions

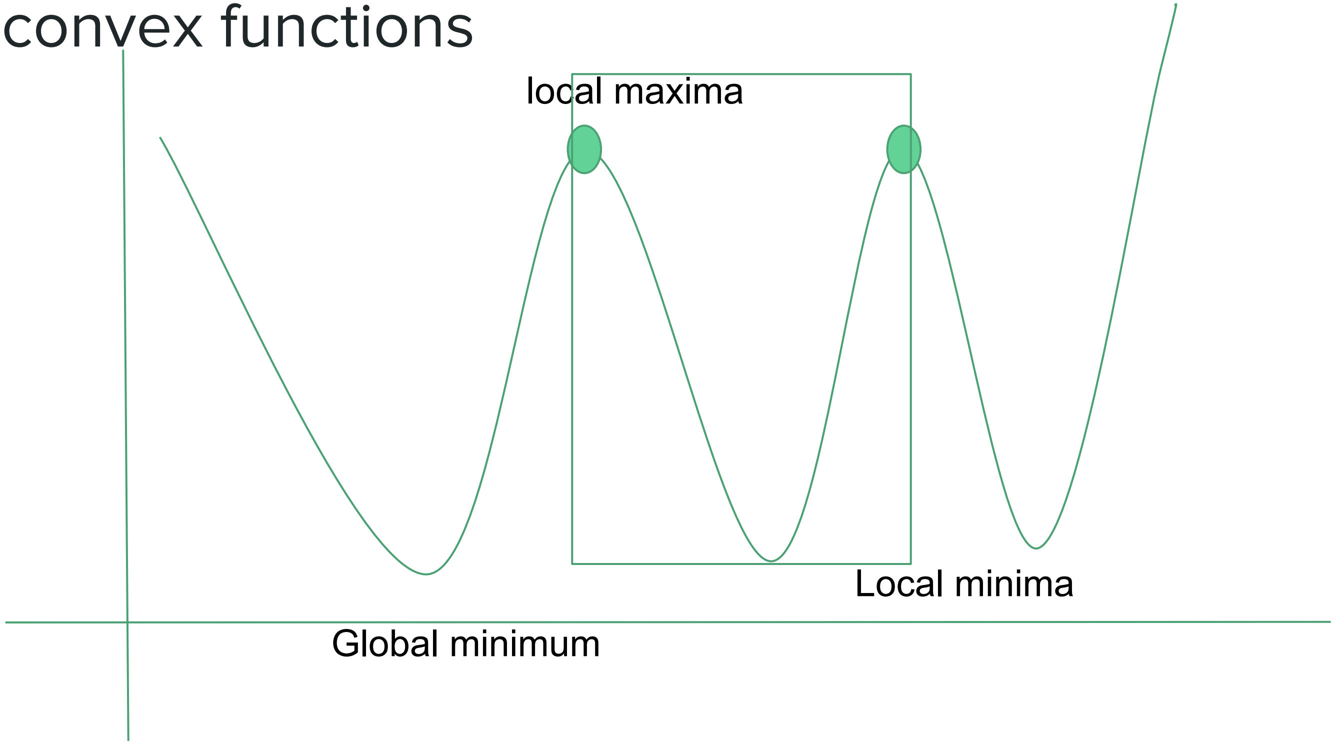


Non-convex functions



Strategy 1: local optimization of the non-convex function

Non-convex functions



Assumptions for local non-convex optimization

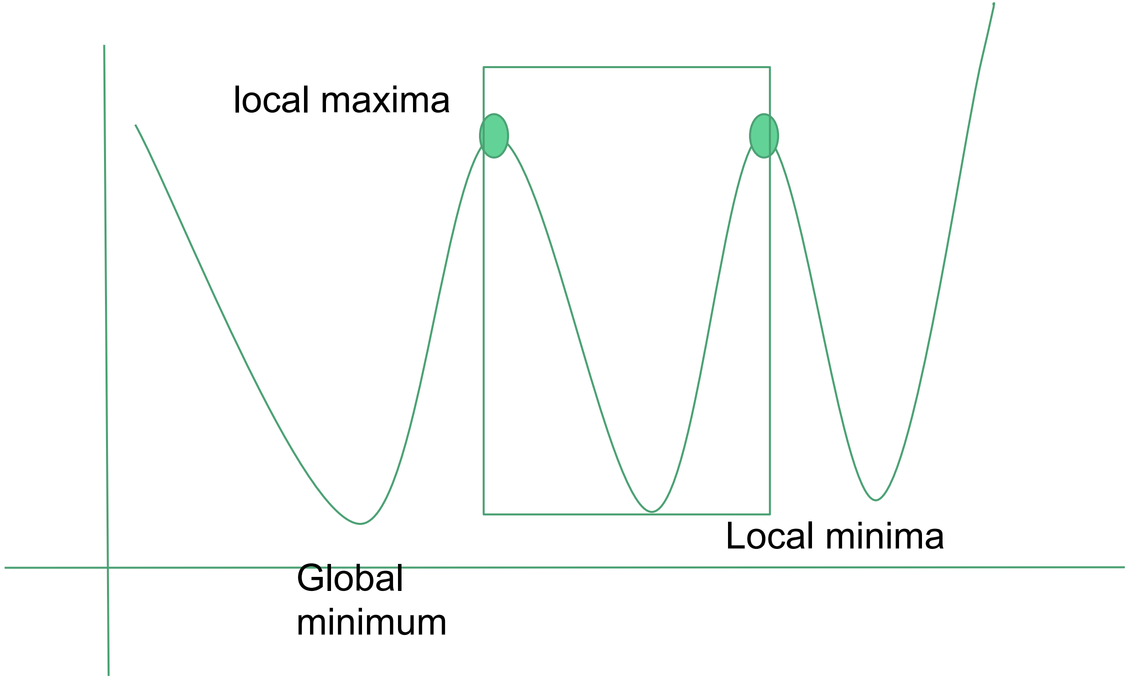
Lipschitz continuous

$$f(y) \leq f(x) + \nabla f(x)(y-x) + \frac{L}{2}(y-x)^2$$

Locally convex

$$||\nabla f(x)||^2 \geq 2\mu(f(x) - f^*).$$

Non-convex functions



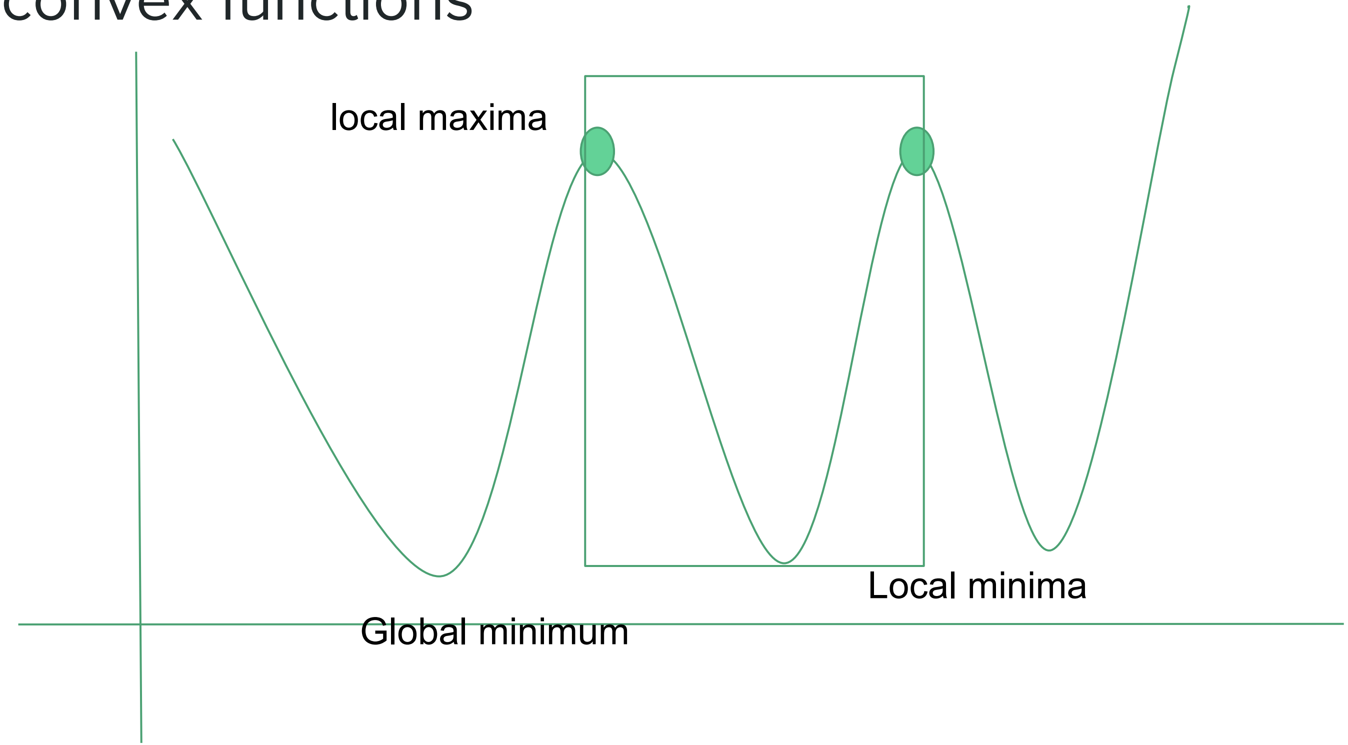
Local randomized coordinate descent

$$E[f(x^{t+1}) - f(x^t)] \leq (1 - \frac{\mu}{nL})[f(x^t) - f(x^*)]$$

Strategy 1: local optimization of the non-convex function

All convex functions rates apply.

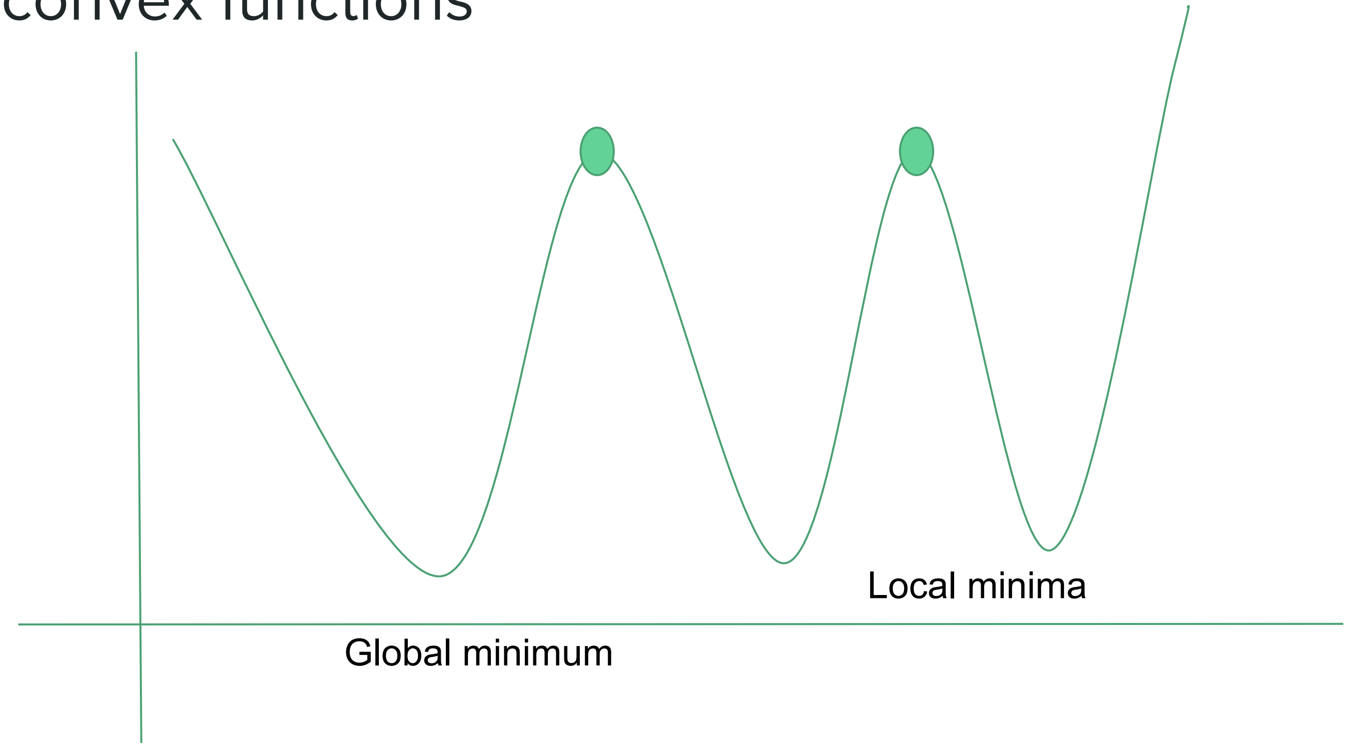
Non-convex functions



Strategy 1: local optimization of the non-convex function

Issue: dealing with saddle points

Non-convex functions



Strategy 2: **Global optimization** of the non-convex function

Issue: Exponential number of saddle points

Local non-convex optimization

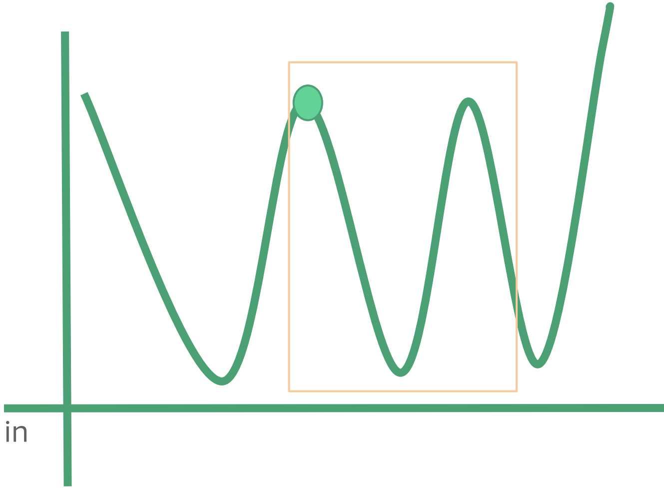
- **Gradient Descent**
  - Difficult to define a proper step size

$$x^{t+1} = x^t - \alpha \nabla f(x^t)$$

- **Newton method**
  - Newton method solves the slowness problem by rescaling the gradients in each direction with the inverse of the corresponding eigenvalues of the hessian
  - can result in moving in the wrong direction (negative eigenvalues)

$$x^{t+1} = x^t - \nabla f(x^t) \nabla^2 f(x^t)^{-1}$$

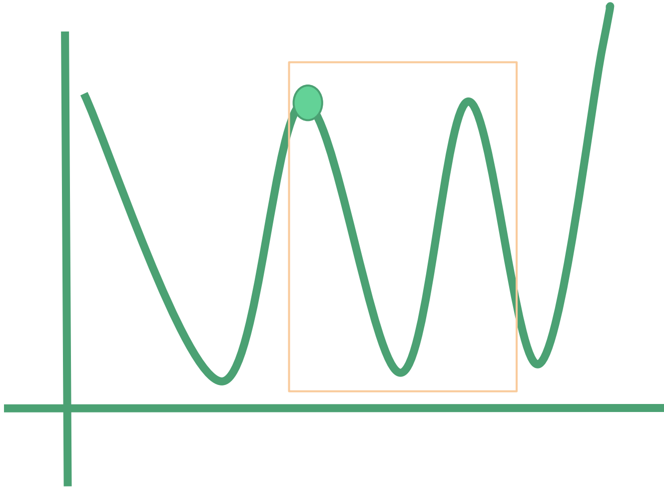
- **Saddle-Free Newton's method**
  - rescales gradients by the absolute value of the inverse Hessian and the Hessian's Lanczos vectors.



Local non-convex optimization

- **Random stochastic gradient descent**
  - Sample noise *r* uniformly from unit sphere
  - Escapes saddle points but step size is difficult to determine

$$x^{t+1} = x^t - \alpha[\nabla f(x^t) + r]$$



- **Cubic regularization [Nesterov 2006]**

Gradient Lipschitz continuous

$$f(y) \le f(x) + \nabla f(x)(y-x) + \tfrac{L}{2}(y-x)^2$$

Hessian Lipschitz continuous

$$f(x^{t+1}) \le f(x^t) + \nabla f(x^t)(x^{t+1} - x^t) + \frac{\nabla^2 f(x^t)}{2}(x^{t+1} - x^t)^2 + \frac{M}{6}|x^{t+1} - x^t|^3$$

Then there exist constants $\epsilon, \delta > 0$ such that whenever a point x_i appears to be in a set $Q = \{x : \|x - \bar{x}\| \le \epsilon, f(x) \ge f(\bar{x})\}$ (for instance, if $x_i = \bar{x}$), then the next point x_{i+1} leaves the set Q:

$$f(x_{i+1}) \le f(\bar{x}) - \delta.$$

Local non-convex optimization

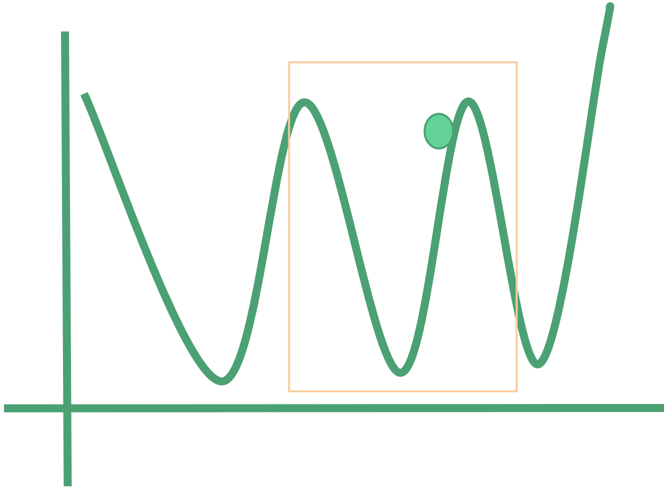
- **Random stochastic gradient descent**
  - Sample noise *r* uniformly from unit sphere
  - Escapes saddle points but step size is difficult to determine

$$x^{t+1} = x^t - \alpha[\nabla f(x^t) + r]$$

- **Momentum**
  - can help escape saddle points (rolling ball)

$$v^{t+1} = \rho v^t - \alpha \nabla f(x^t)$$

$$x^{t+1} = x^t + v^{t+1}$$



Global non-convex optimization

- Matrix completion problem [De Sa et al. 2015]

$$
\begin{aligned}
&\textbf{minimize} \quad \mathbf{E}\left[\left\|\tilde{A} - X\right\|_F^2\right] \\
&\textbf{subject to} \quad X \in \mathbb{R}^{n\times n}, \mathbf{rank}\,(X) \le p, X \succeq 0,
\end{aligned}
$$

- Applications (usually for datasets with missing data)
  - matrix completion
  - Image reconstruction
  - recommendation systems.

Global non-convex optimization

- Matrix completion problem [De Sa et al. 2015]

$$\begin{aligned} &\textbf{minimize} \quad \mathbf{E}\left[\|\tilde{A} - X\|_F^2\right] \\ &\textbf{subject to} \quad X \in \mathbb{R}^{n\times n}, \textbf{rank}\,(X) \le p, X \succeq 0, \end{aligned}$$

- Reformulate it as (unconstrained non-convex problem)

$$\begin{aligned} &\textbf{minimize} \quad \mathbf{E}\left[\|\tilde{A} - YY^T\|_F^2\right] \\ &\textbf{subject to} \quad Y \in \mathbb{R}^{n\times p} \end{aligned}.$$

- Gradient descent

$$Y_{k+1} = Y_k + \alpha_k \left(\tilde{A}_k - Y_k Y_k^T\right) Y_k.$$

Global non-convex optimization

- Reformulate it as (unconstrained non-convex problem)

$$\begin{array}{ll} \text{minimize} & \mathbf{E}\left[\|\tilde{A} - YY^T\|_F^2\right] \\ \text{subject to} & Y \in \mathbb{R}^{n \times p} \end{array}.$$

- Gradient descent

$$Y_{k+1} = Y_k + \alpha_k \left(\tilde{A}_k - Y_k Y_k^T\right) Y_k.$$

- Using the Riemannian manifold, we can derive the following

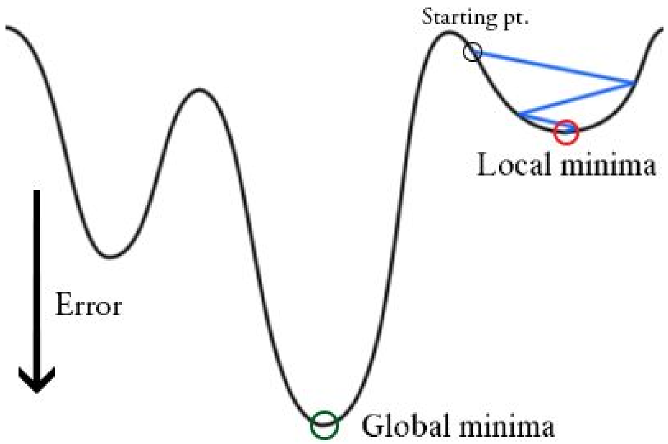
$$Y_{k+1} = \left(I + \eta_k \tilde{A}_k\right) Y_k.$$

- This widely-used algorithm converges globally, using only random initialization

Convex relaxation of non-convex functions optimization

- Convex Neural Networks [Bengio et al. 2006]
  - Single-hidden layer network

$$\boxed{L(x,y) = \frac{1}{2}||f(x,y) - b||^2} \qquad f(x,y) = y \cdot g(Ax)$$



original neural networks non-convex problem

Convex relaxation of non-convex functions optimization

- Convex Neural Networks [Bengio et al. 2006]
  - Single-hidden layer network

$$\boxed{L(x, y) = \frac{1}{2}||f(x, y) - b||^2}$$

$$f(x, y) = y \cdot g(Ax)$$

  - Use alternating minimization

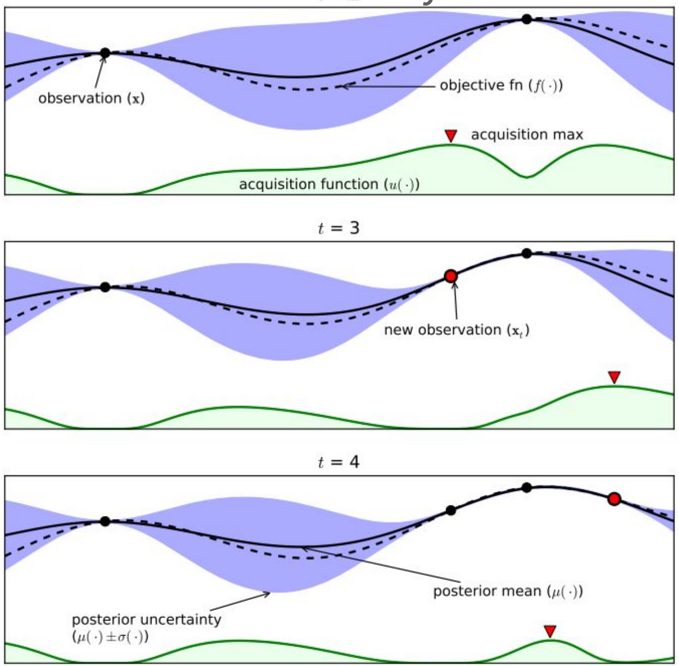
$$L_c(x) = \frac{1}{2}||f_c(x) - b||^2 \qquad f_c(x) = h(v) \cdot x$$

$$L_h(v) = \frac{1}{2}||h(v) - f_c(x)||^2 \qquad h(v) = g(Av)$$

  - Potential issues with the activation function

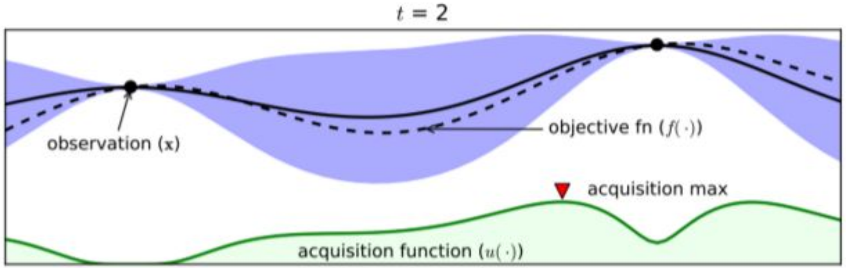
Bayesian optimization (global non-convex optimization)

- Typically used for finding optimal parameters
  - Determining the step size of # hidden layers for neural networks
  - The parameter values are bounded ?
- Other methods include sampling the parameter values random uniformly
  - Grid-search



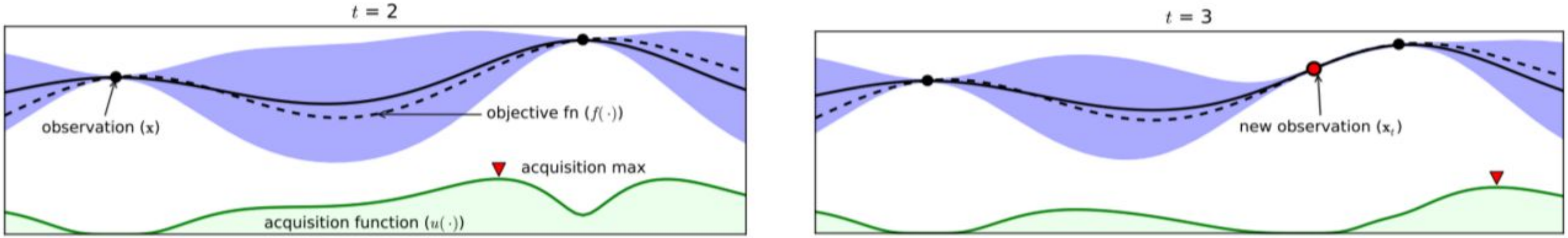
Bayesian optimization (global non-convex optimization)

- Fit Gaussian process on the observed data (purple shade)
  - Probability distribution on the function values



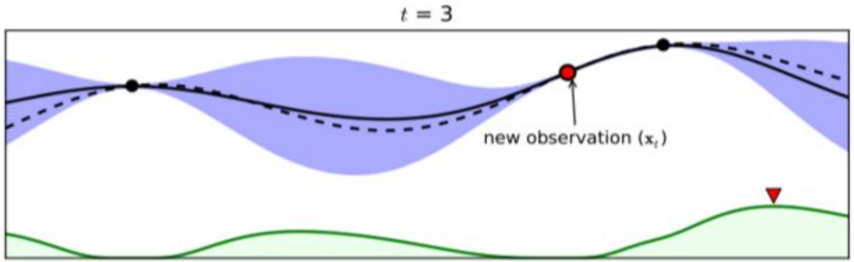
Bayesian optimization (global non-convex optimization)

- Fit Gaussian process on the observed data (purple shade)
  - Probability distribution on the function values
- Acquisition function (green shade)
  - a function of
    - the objective value (exploitation) in the Gaussian density function; and
    - the uncertainty in the prediction value (exploration).



Bayesian optimization

- Slower than grid-search with low level of smoothness (illustrate)
- Faster than grid-search with high level of smoothness (illustrate)



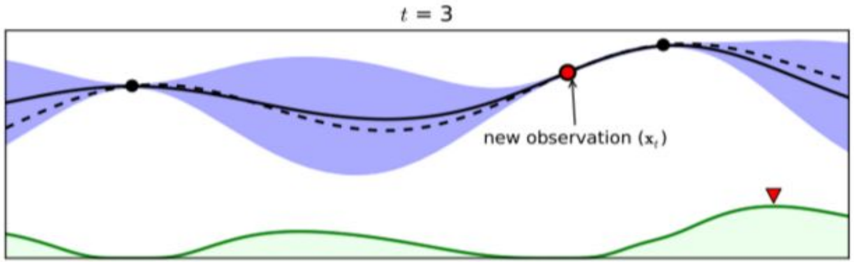
Bayesian optimization

- Slower than grid-search with low level of smoothness (illustrate)
- Faster than grid-search with high level of smoothness (illustrate)

Improves error from $O(1/t^{1/d})$ to $O(1/t^{\nu/d})$

Grid-search

Bayesian optimization



Bayesian optimization

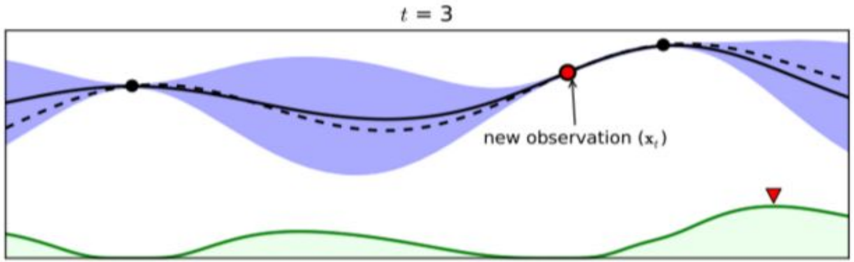
- Slower than grid-search with low level of smoothness (illustrate)
- Faster than grid-search with high level of smoothness (illustrate)

A measure of smoothness

Improves error from $O(1/t^{1/d})$ to $O(1/t^{\boxed{v}/d})$

Grid-search

Bayesian optimization



Summary

Non-convex optimization

- Strategy 1: Local non-convex optimization
  - Convexity convergence rates apply
  - Escape saddle points using, for example, cubic regularization and saddle-free newton update
- Strategy 2: Relaxing the non-convex problem to a convex problem
  - Convex neural networks
- Strategy 3: Global non-convex optimization
  - Bayesian optimization
  - Matrix completion