Introduction to stochastic optimization

Dmitry A. Kropotov

Gradient descent

Optimization problem:

$$F(\boldsymbol{x}) \to \min_{\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^d, \; F \text{ – some smooth function.}$$

Gradient descent

Optimization problem: 

$$F(\boldsymbol{x}) \to \min_{\boldsymbol{x}}, \quad \boldsymbol{x} \in \mathbb{R}^d, \; F \text{ – some smooth function.}$$

Gradient descent method:

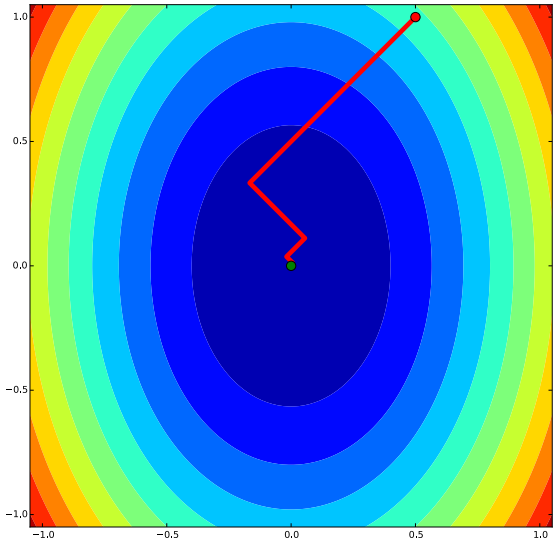
- Iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k)$;
- Adapt stepsize α_k using some function decreasing criterion, e.g. Armijo rule:

$$\alpha_k : \; F(\boldsymbol{x}_{k+1}) \leq F(\boldsymbol{x}_k) - \alpha_k c_1 \|\nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k)\|^2, \; c_1 = 10^{-4};$$

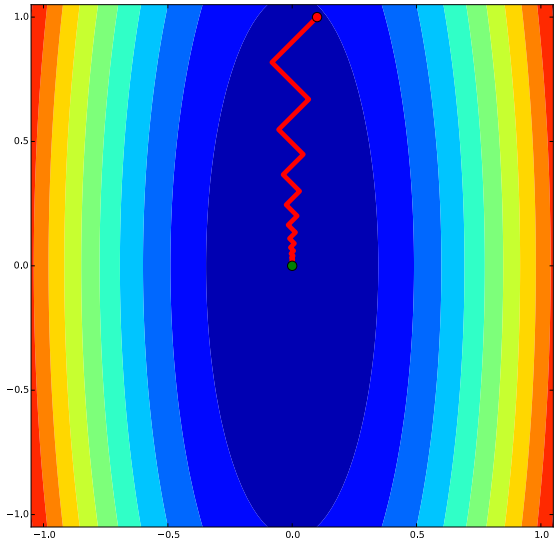
- Stop when $\|\nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k)\|^2 \leq \varepsilon$.

Trajectories of Gradient Descent

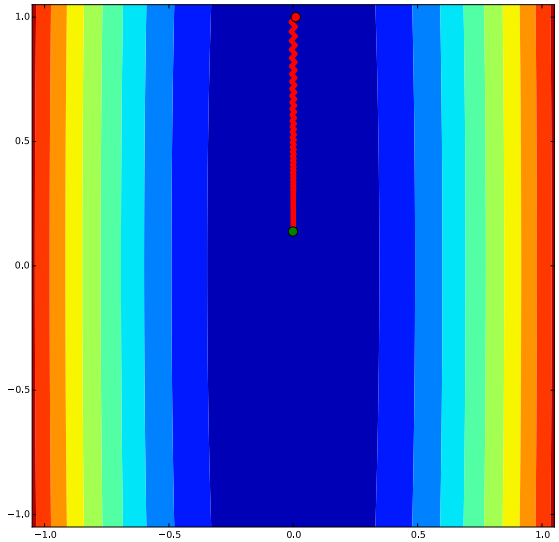
$$F(x, y) = \frac{1}{2}x^2 + \frac{\rho}{2}y^2 \to \min_{x,y}$$



$\rho = 0.5$



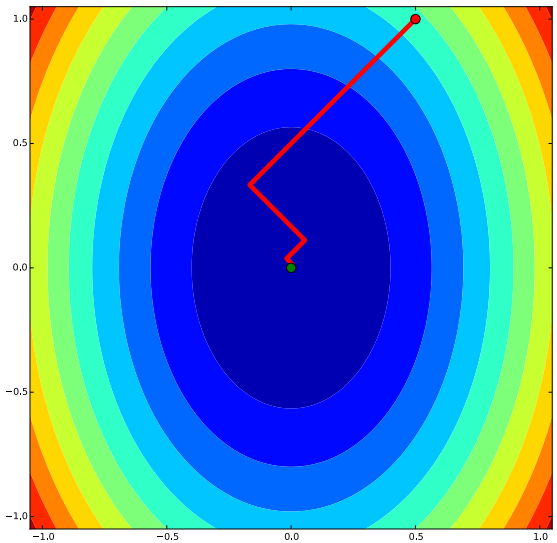
$\rho = 0.1$



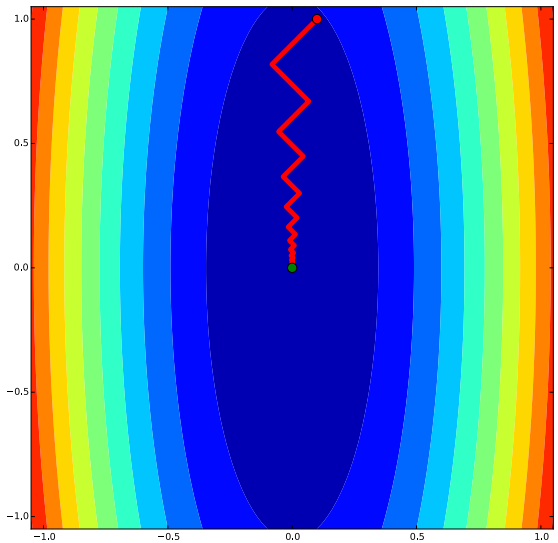
$\rho = 0.01$

Trajectories of Gradient Descent

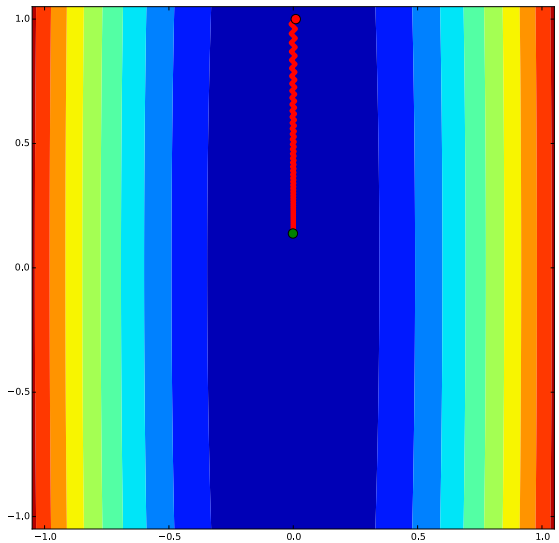
$$F(x, y) = \frac{1}{2}x^2 + \frac{\rho}{2}y^2 \to \min_{x,y}$$



$\rho = 0.5$



$\rho = 0.1$



$\rho = 0.01$

Gradient descent is sensitive to poor function scaling!

Newton method

$$F(\boldsymbol{x}) \approx m_k(\boldsymbol{x}) = F(\boldsymbol{x}_k) + \boldsymbol{g}_k^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^T H_k(\boldsymbol{x} - \boldsymbol{x}_k) \to \min_{\boldsymbol{x}}$$

$$\boldsymbol{g}_k = \nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k), \;\; H_k = \nabla^2_{\boldsymbol{x}} F(\boldsymbol{x}_k),$$

$$\arg\min_{\boldsymbol{x}} m_k(\boldsymbol{x}) = \boldsymbol{x}_k - H_k^{-1}\boldsymbol{g}_k.$$

Newton method

$$F(\boldsymbol{x}) \approx m_k(\boldsymbol{x}) = F(\boldsymbol{x}_k) + \boldsymbol{g}_k^T(\boldsymbol{x} - \boldsymbol{x}_k) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_k)^T H_k(\boldsymbol{x} - \boldsymbol{x}_k) \to \min_{\boldsymbol{x}}$$

$$\boldsymbol{g}_k = \nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k), \;\; H_k = \nabla^2_{\boldsymbol{x}} F(\boldsymbol{x}_k),$$

$$\arg\min_{\boldsymbol{x}} m_k(\boldsymbol{x}) = \boldsymbol{x}_k - H_k^{-1}\boldsymbol{g}_k.$$

Newton method:

- Iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k H_k^{-1}\boldsymbol{g}_k$;
- Adapt stepsize α_k using some function decreasing criterion, e.g. Armijo rule:

$$\alpha_k : \;\; F(\boldsymbol{x}_{k+1}) \leq F(\boldsymbol{x}_k) - \alpha_k c_1 \boldsymbol{g}_k^T H_k^{-1}\boldsymbol{g}_k, \;\; c_1 = 10^{-4};$$

- Stop when $\|\nabla_{\boldsymbol{x}} F(\boldsymbol{x}_k)\|^2 \leq \varepsilon$.

H_k^{-1} gives proper gradient scaling!

Advanced optimization methods

Limitations of Newton method:

- Hard to calculate Hessian H_k on each iteration;
- Hard to store Hessian in memory;
- Hard to inverse Hessian;
- Needs some Hessian modifications in case of non-convex function F.

Advanced optimization methods

Limitations of Newton method:

- Hard to calculate Hessian H_k on each iteration;
- Hard to store Hessian in memory;
- Hard to inverse Hessian;
- Needs some Hessian modifications in case of non-convex function F.

Advanced first-order optimization strategies:

- Conjugate gradients;
- Hessian-free Newton;
- Quasi-Newton methods (e.g. L-BFGS);
- etc.

Stochastic optimization problem

General stochastic optimization problem:

$$F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}) \to \min_{\boldsymbol{x}}, \; q - \text{some probability distribution.}$$

Stochastic optimization problem

General stochastic optimization problem:

$$F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}) \to \min_{\boldsymbol{x}}, \; q - \text{some probability distribution.}$$

Examples:

- Empirical risk minimization:

$$F(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{x}) = \mathbb{E}_{i \sim \text{Unif}\{1,\ldots,N\}} f_i(\boldsymbol{x}) \to \min_{\boldsymbol{x}},$$

f_i – loss function for i-th training object.

Stochastic optimization problem

General stochastic optimization problem:

$$F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}) \to \min_{\boldsymbol{x}}, \;\; q \text{ – some probability distribution.}$$

Examples:

- Empirical risk minimization:

$$F(\boldsymbol{x}) = \frac{1}{N} \sum_{i=1}^{N} f_i(\boldsymbol{x}) = \mathbb{E}_{i \sim \text{Unif}\{1,\ldots,N\}} f_i(\boldsymbol{x}) \to \min_{\boldsymbol{x}},$$

f_i – loss function for i-th training object.

- Evidence lower bound (ELBO) maximization:

$$\mathbb{E}_{q(\boldsymbol{t}|\boldsymbol{\lambda})} \log p(\boldsymbol{x}, \boldsymbol{t}|\boldsymbol{\theta}) - \mathbb{E}_{q(\boldsymbol{t}|\boldsymbol{\lambda})} \log q(\boldsymbol{t}|\boldsymbol{\lambda}) \to \max_{\boldsymbol{\theta}, \boldsymbol{\lambda}}.$$

Computational challenges

$$F(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x}) \to \min_{\boldsymbol{x}}.$$

Value	Calculation costs:
$f_i(\boldsymbol{x})$	O(s)
$\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})$	

Computational challenges

$$F(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x}) \to \min_{\boldsymbol{x}}.$$

Value	Calculation costs:
$f_i(\boldsymbol{x})$	O(s)
$\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})$	O(s)
$F(\boldsymbol{x})$	O(Ns)
$\nabla F(\boldsymbol{x})$	O(Ns)

Can be challenging when $N \gg 1$!

Computational challenges

$$F(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N} f_i(\boldsymbol{x}) \to \min_{\boldsymbol{x}}.$$

Value	Calculation costs:
$f_i(\boldsymbol{x})$	O(s)
$\nabla_{\boldsymbol{x}} f_i(\boldsymbol{x})$	O(s)
$F(\boldsymbol{x})$	O(Ns)
$\nabla F(\boldsymbol{x})$	O(Ns)

Can be challenging when $N \gg 1$!

$$F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}) = \int f(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{y}) d\boldsymbol{y} \to \min_{\boldsymbol{x}}.$$

Exact calculation may not exist for complex distribution q!

Stochastic estimates

$$F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}),$$
$$g(\boldsymbol{x}) = \nabla_{\boldsymbol{x}} F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}).$$

Stochastic estimates:

$$\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim q(\boldsymbol{y}),$$
$$\hat{F}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{x}, \boldsymbol{y}_i), \quad \hat{g}(\boldsymbol{x}) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}, \boldsymbol{y}_i).$$

Estimates $\hat{F}$ and $\hat{g}$:

- unbiased, i.e. $\mathbb{E}_{q(\boldsymbol{y})} \hat{F}(\boldsymbol{x}) = F(\boldsymbol{x})$, $\mathbb{E}_{q(\boldsymbol{y})} \hat{g}(\boldsymbol{x}) = g(\boldsymbol{x})$;
- become exact when $m \to +\infty$;
- easy to calculate for small m.

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim q(\boldsymbol{y})$;

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim q(\boldsymbol{y})$;
- Calculate $\hat{g}(\boldsymbol{x}_k) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_i)$;

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim q(\boldsymbol{y})$;
- Calculate $\hat{g}(\boldsymbol{x}_k) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_i)$;
- Make a step: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \hat{g}(\boldsymbol{x}_k)$;

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim q(\boldsymbol{y})$;
- Calculate $\hat{g}(\boldsymbol{x}_k) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_i)$;
- Make a step: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \hat{g}(\boldsymbol{x}_k)$;
- Iterate until convergence.

Stochastic Gradient Descent

Using stochastic estimates in gradient descent method leads to Stochastic Gradient Descent (SGD):

- Generate a small batch of samples: $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m \sim q(\boldsymbol{y})$;
- Calculate $\hat{g}(\boldsymbol{x}_k) = \frac{1}{m} \sum_{i=1}^{m} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_k, \boldsymbol{y}_i)$;
- Make a step: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \hat{g}(\boldsymbol{x}_k)$;
- Iterate until convergence.

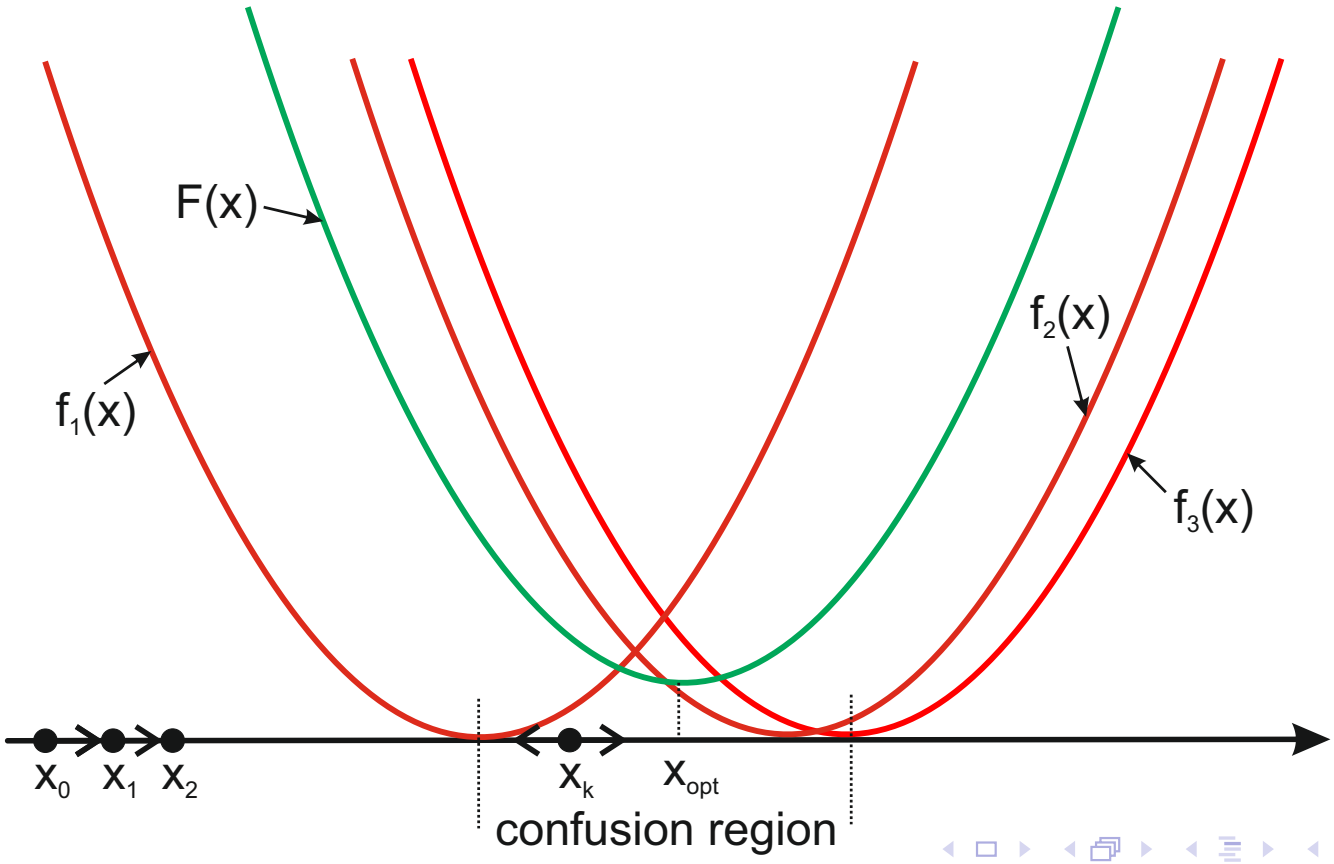
Questions for SGD:

- Convergence rate comparing to full gradient descent?
- How to choose step size α_k?
- When to stop?

SGD trajectory

Fitting linear regression in 1D space:

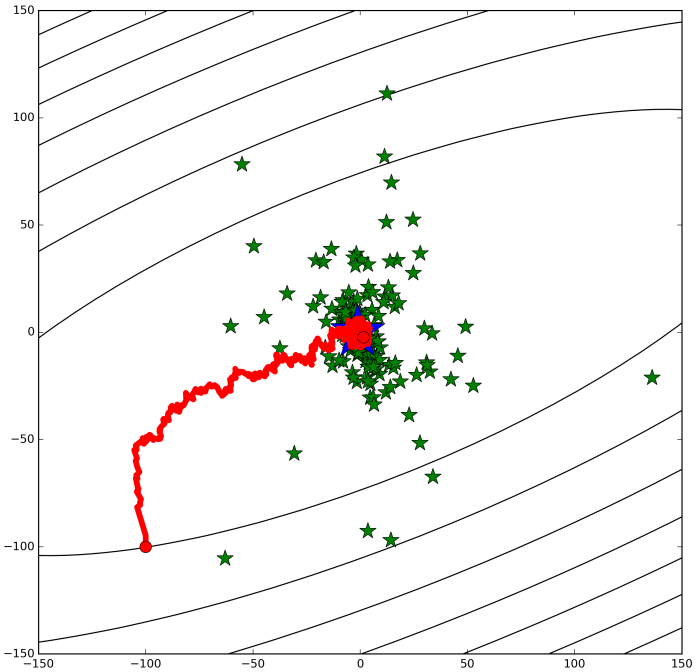
$$F(x) = \frac{1}{N}\sum_{i=1}^{N} \underbrace{\frac{1}{2}(b_i - a_i x)^2}_{f_i(x)} \to \min_{x \in \mathbb{R}}.$$



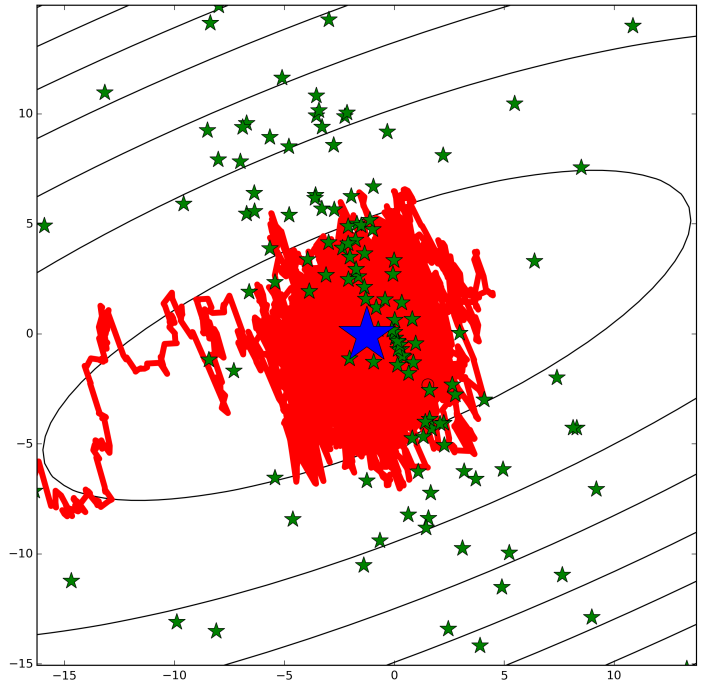
SGD trajectory

Fitting linear regression in 2D space:

$$F(\boldsymbol{x}) = \frac{1}{N}\sum_{i=1}^{N}(b_i - \boldsymbol{a}_i^T\boldsymbol{x})^2 \to \min_{\boldsymbol{x}\in\mathbb{R}^2}.$$



SGD trajectory



optimum vicinity

SGD properties

SGD properties:

- stochastic gradient doesn't generate descent direction;
- stochastic gradient doesn't equal zero in optimum of $F(\boldsymbol{x}) \Rightarrow$ SGD can't converge with constant step size;
- we don't have access to $F(\boldsymbol{x}) \Rightarrow$ can't use Armijo-like rules for adapting step size;
- we don't have access to $\nabla_{\boldsymbol{x}} F(\boldsymbol{x}) \Rightarrow$ can't use standard termination condition.

SGD convergence analysis

Optimization problem: $F(\boldsymbol{x}) = \mathbb{E}_{q(\boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y}) \to \min_{\boldsymbol{x}}$

SGD iteration: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \hat{\boldsymbol{g}}_k$, $\mathbb{E}\hat{\boldsymbol{g}}_k = \boldsymbol{g}_k = \nabla F(\boldsymbol{x}_k)$.

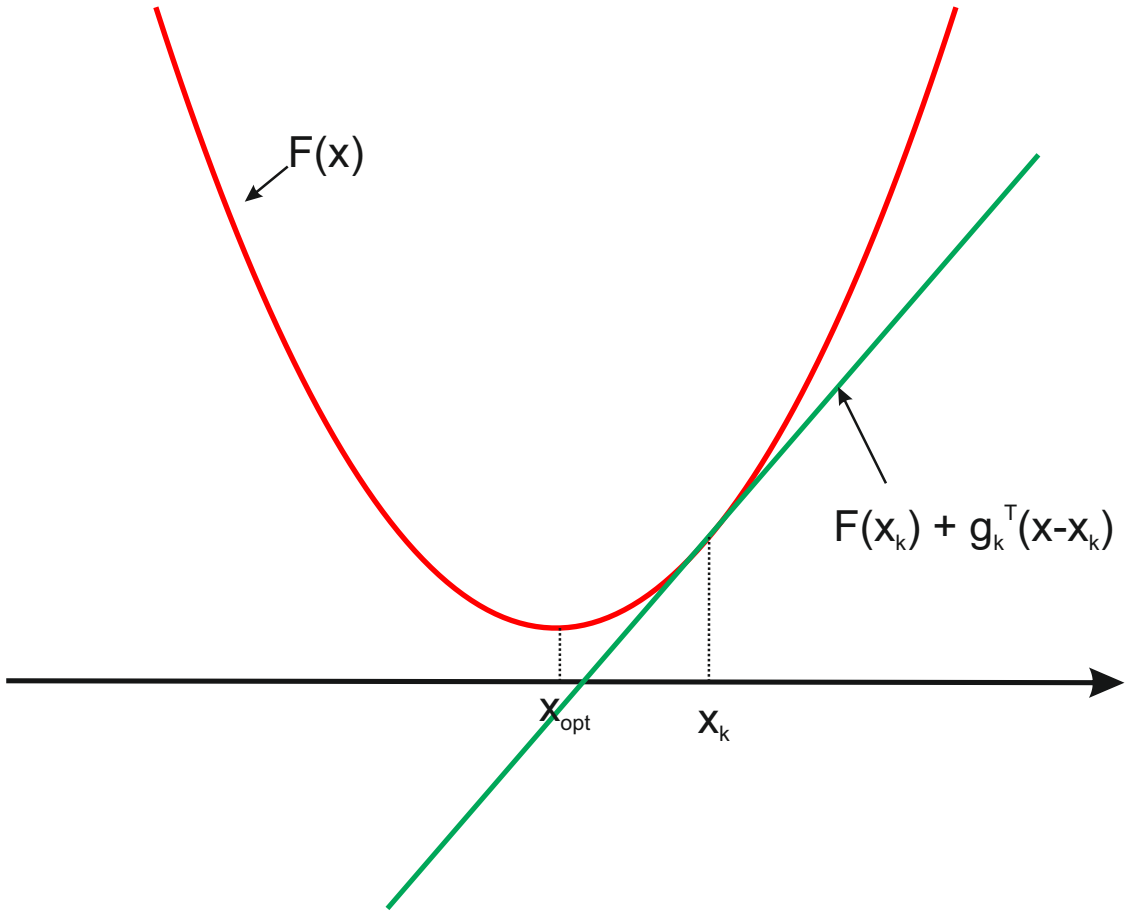
$$\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_{opt}\|^2 = \|\boldsymbol{x}_k - \boldsymbol{x}_{opt} - \alpha_k \hat{\boldsymbol{g}}_k\|^2 =$$
$$= \|\boldsymbol{x}_k - \boldsymbol{x}_{opt}\|^2 - 2\alpha_k \hat{\boldsymbol{g}}_k^T (\boldsymbol{x}_k - \boldsymbol{x}_{opt}) + \alpha_k^2 \|\hat{\boldsymbol{g}}_k\|^2$$

$$\mathbb{E}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_{opt}\|^2 = \|\boldsymbol{x}_k - \boldsymbol{x}_{opt}\|^2 - 2\alpha_k \boldsymbol{g}_k^T (\boldsymbol{x}_k - \boldsymbol{x}_{opt}) + \alpha_k^2 \mathbb{E}\|\hat{\boldsymbol{g}}_k\|^2$$

SGD convergence analysis

Suppose F is a convex function, i.e.

$$F(\boldsymbol{x}_{opt}) \geq F(\boldsymbol{x}_k) + \boldsymbol{g}_k^T(\boldsymbol{x}_{opt} - \boldsymbol{x}_k)$$



SGD convergence analysis

Таким образом имеем:

$$\mathbb{E}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_{opt}\|^2 = \|\boldsymbol{x}_k - \boldsymbol{x}_{opt}\|^2 - 2\alpha_k \boldsymbol{g}_k^T(\boldsymbol{x}_k - \boldsymbol{x}_{opt}) + \alpha_k^2 \mathbb{E}\|\hat{\boldsymbol{g}}_k\|^2 \quad (1)$$

$$F(\boldsymbol{x}_{opt}) \geq F(\boldsymbol{x}_k) + \boldsymbol{g}_k^T(\boldsymbol{x}_{opt} - \boldsymbol{x}_k). \quad (2)$$

Тогда

$$\alpha_k(F(\boldsymbol{x}_k) - F_{opt}) \overset{(2)}{\leq} \alpha_k \boldsymbol{g}_k^T(\boldsymbol{x}_k - \boldsymbol{x}_{opt}) \overset{(1)}{=} \frac{1}{2}\|\boldsymbol{x}_k - \boldsymbol{x}_{opt}\|^2 + \\ + \frac{1}{2}\alpha_k^2 \mathbb{E}\|\hat{\boldsymbol{g}}_k\|^2 - \frac{1}{2}\mathbb{E}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_{opt}\|^2.$$

Берём мат. ожидание от левой части и суммируем по k от нуля:

$$\sum_{i=0}^{k} \alpha_i(\mathbb{E}F(\boldsymbol{x}_i) - F_{opt}) \leq \frac{1}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_{opt}\|^2 + \frac{1}{2}\sum_{i=0}^{k} \alpha_i^2 \mathbb{E}\|\hat{\boldsymbol{g}}_i\|^2 - \\ - \frac{1}{2}\mathbb{E}\|\boldsymbol{x}_{k+1} - \boldsymbol{x}_{opt}\|^2 \leq \frac{1}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_{opt}\|^2 + \frac{1}{2}\sum_{i=0}^{k} \alpha_i^2 \mathbb{E}\|\hat{\boldsymbol{g}}_i\|^2.$$

SGD convergence analysis

$$\mathbb{E}F\Bigg(\underbrace{\frac{\sum_{i=0}^{k}\alpha_i \boldsymbol{x}_i}{\sum_{i=0}^{k}\alpha_i}}_{\hat{\boldsymbol{x}}_k}\Bigg) - F_{opt} \leq \{\text{convexity of F}\} \leq$$

$$\frac{\sum_{i=0}^{k}\alpha_i(\mathbb{E}F(\boldsymbol{x}_i) - F_{opt})}{\sum_{i=0}^{k}\alpha_i} \leq \frac{\frac{1}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}_{opt}\|^2 + \frac{1}{2}\sum_{i=0}^{k}\alpha_i^2 \mathbb{E}\|\hat{\boldsymbol{g}}_i\|^2}{\sum_{i=0}^{k}\alpha_i}.$$

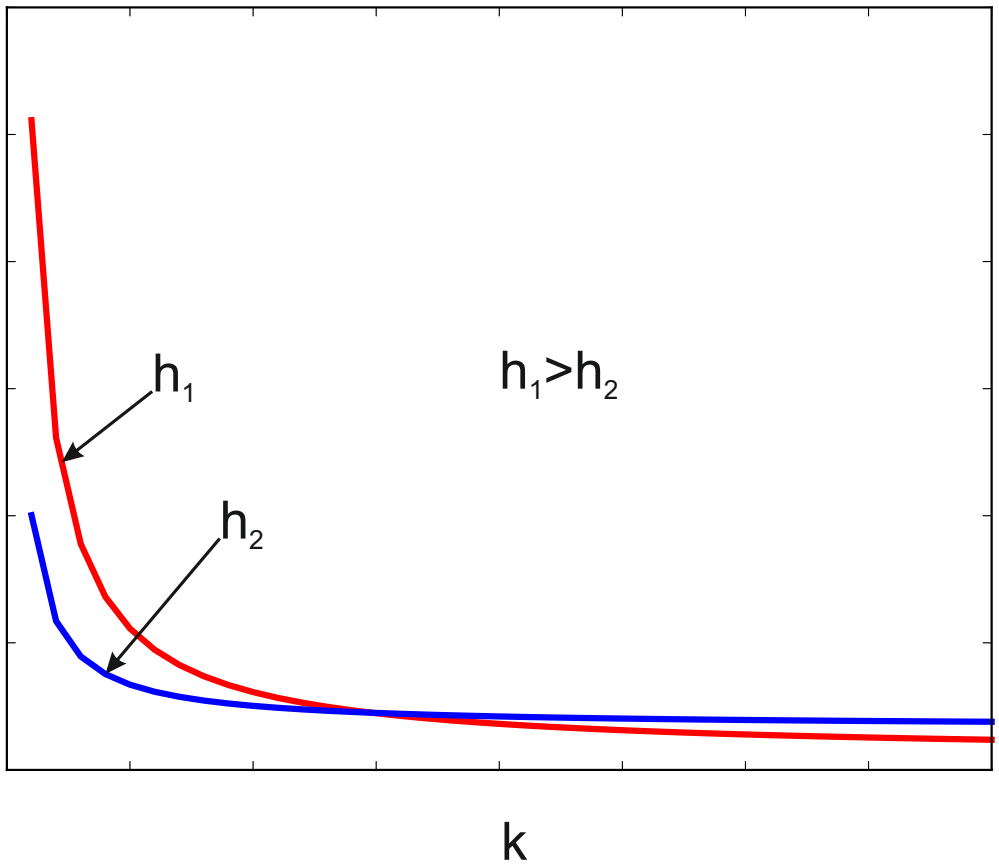
Suppose $\|\boldsymbol{x}_0 - \boldsymbol{x}_{opt}\|^2 \leq R^2$ and $\mathbb{E}\|\hat{\boldsymbol{g}}_k\|^2 \leq G^2$. Then

$$\boxed{\mathbb{E}F(\hat{\boldsymbol{x}}_k) - F_{opt} \leq \frac{R^2 + G^2\sum_{i=0}^{k}\alpha_i^2}{2\sum_{i=0}^{k}\alpha_i}}$$

Choosing step size

Strategy 1: $\alpha_i = h \ \forall i$. Then

$$\mathbb{E}F(\hat{\boldsymbol{x}}_k) - F_{opt} \leq \frac{R^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i} = \frac{R^2}{2h(k+1)} + \frac{G^2 h}{2} \underset{k \to \infty}{\longrightarrow} \frac{G^2 h}{2}.$$



Choosing step size

$$\boxed{\mathbb{E}F(\hat{\boldsymbol{x}}_k) - F_{opt} \leq \frac{R^2 + G^2 \sum_{i=0}^{k} \alpha_i^2}{2 \sum_{i=0}^{k} \alpha_i}}$$

Strategy 2: $\sum_{i=0}^{\infty} \alpha_i = \infty$, $\sum_{i=0}^{\infty} \alpha^2 < \infty$. This is true for the following choice:

$$\alpha_i = \frac{\eta}{(i+1)^\tau}, \ \tau \in (1/2, 1].$$

If $\tau = 1$, then $\sum_{i=0}^{k} \frac{1}{i} \sim \log k$ and $\mathbb{E}F(\hat{\boldsymbol{x}}_k) - F_{opt} = O(1/\log k)$.

Strategy 3: $\sum_{i=0}^{\infty} \alpha_i = \infty$, $\sum_{i=0}^{k} \alpha^2 / \sum_{i=0}^{k} \alpha_i \underset{k \to \infty}{\longrightarrow} 0$. This is true for the following choice:

$$\alpha_i = \frac{\eta}{(i+1)^\tau}, \ \tau \leq 1/2.$$

If $\tau = 1/2$, then $\mathbb{E}F(\hat{\boldsymbol{x}}_k) - F_{opt} = O(\log k / \sqrt{k})$. This is optimal choice.

GD and SGD convergence rates

Function F	GD	SGD
Smooth and strongly convex	$O(c^k)$	$O(1/k)$
Smooth and convex	$O(1/k)$	$O(1/\sqrt{k})$

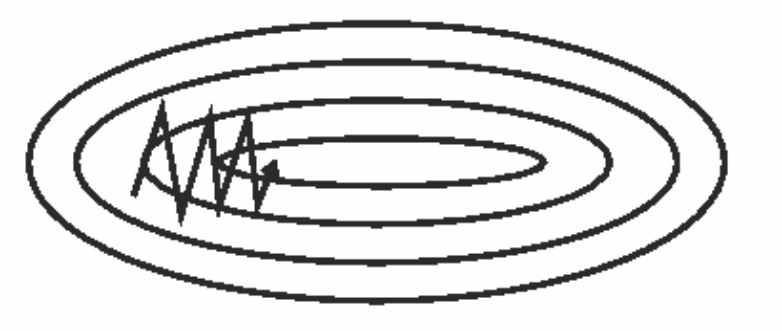
Notes on SGD

- Full convergence is very slow $\Rightarrow$ we can't expect solution with high accuracy;
- Valid stopping criterion is absent $\Rightarrow$ run the method for appropriate time period and measure the quality of solution using testing or control sample;
- In practice it is usually enough to reach confusion region, so choose strategy 1 for step size and possibly decrease the step size when testing quality stabilizes;
- It is important to have low G.

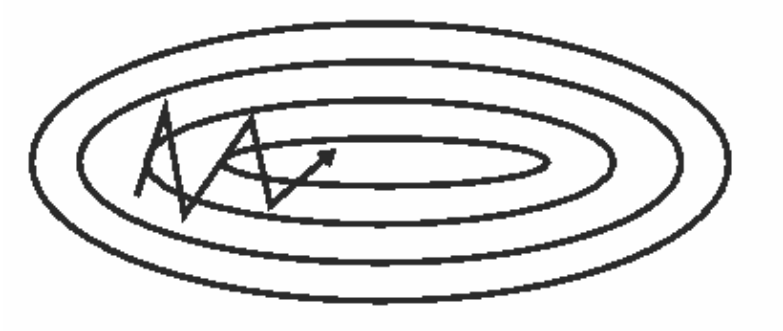
Using SGD with momentum

SGD with momentum:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \hat{\boldsymbol{g}}_k + \beta_k(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}).$$

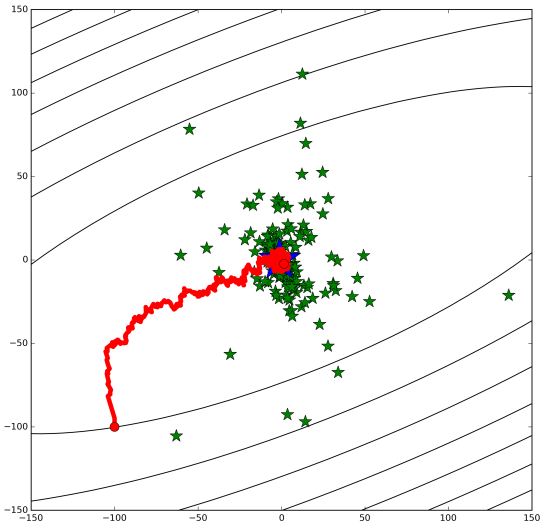


without momentum

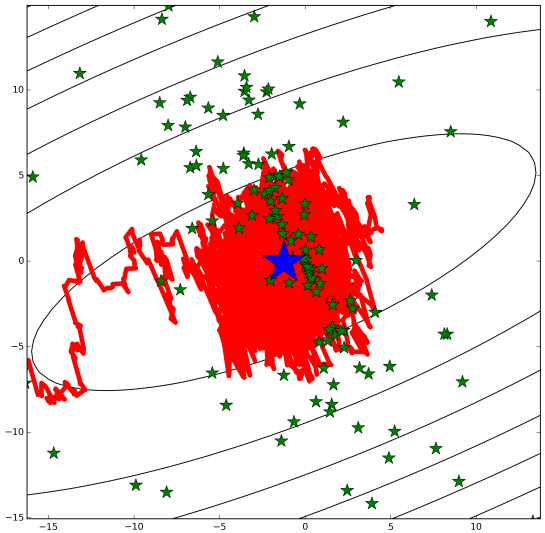


with momentum

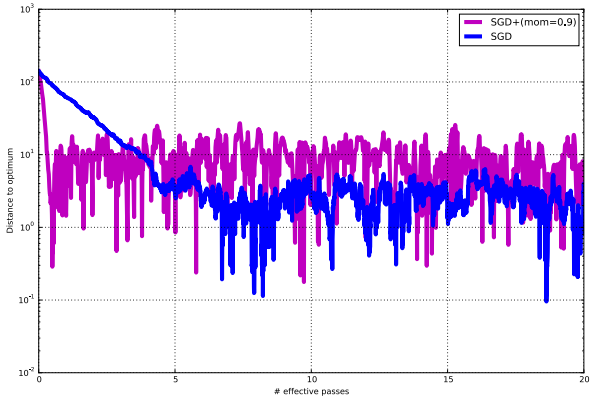
Using SGD with momentum

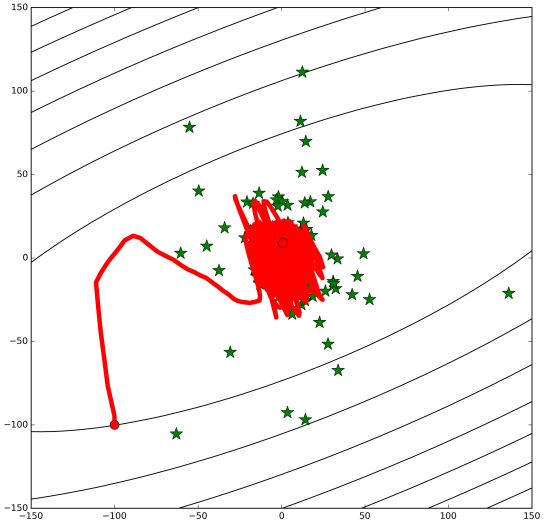


SGD

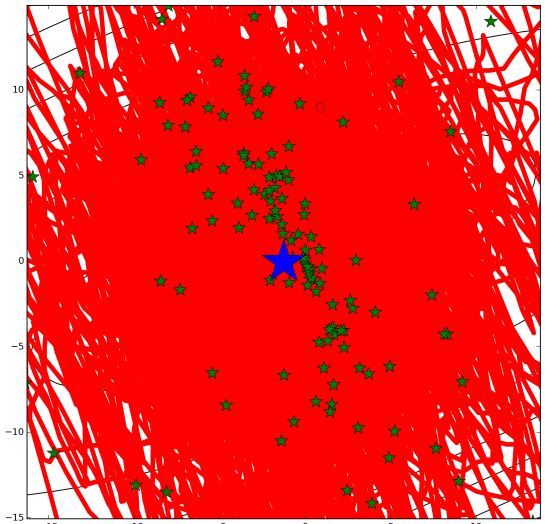


SGD in opt. vicinity





SGD+mom.



SGD+mom. in opt. vicinity

Adam

$$\boldsymbol{m}_{k+1} = \beta_1 \boldsymbol{m}_k + (1 - \beta_1)\hat{\boldsymbol{g}}_k,$$
$$\boldsymbol{v}_{k+1} = \beta_2 \boldsymbol{v}_k + (1 - \beta_2)\hat{\boldsymbol{g}}_k \odot \hat{\boldsymbol{g}}_k,$$
$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \alpha_k \frac{\boldsymbol{m}_{k+1}}{\sqrt{\boldsymbol{v}_{k+1} + \varepsilon}}.$$