Statsitical Analysis of On-line Learning

optimal and semi-optimal stochastic gradient

Noboru Murata

June 12, 2023

https://noboru-murata.github.io/

Introduction

notation:

- data: i.i.d.~ observations from ground truth distribution P

$$z_1, z_2, \dots, z_t, \dots \sim^{\text{i.i.d.}} P$$

- learning machine: specified by a finite dimensional parameter

$$\theta \in \Theta \subset \mathbb{R}^m$$

- loss function: penalty of machine θ for a given datum z

$$l(z;\theta) \quad \text{(a smooth function with respect to } \theta\text{)}$$

for example

$$l(z;\theta) = -\log p(z:\theta) \qquad \text{negative log loss}$$
$$l(z;\theta) = |y - f(x;\theta)|^2 \qquad \text{squared loss for } z = (x, y)$$

- population loss: not accessible

$$L(\theta) = \mathbb{E}_{Z\sim P}[l(Z;\theta)]$$

$$\theta = \arg\min_{\theta} L(\theta) \quad \text{(optimal parameter)}$$

- empirical loss: accessible

$$\hat{L}_t(\theta) = \frac{1}{t}\sum_{z_i\in D_t} l(z_i;\theta), \quad D_t = \{z_i; i = 1,\ldots,t\}$$

- $\hat{L}$ is justified by *the law of large numbers*

$$\hat{L}_t(\theta) = \frac{1}{t}\sum_{z_i\in D_t} l(z_i;\theta) \xrightarrow{t\to\infty} L(\theta) = \mathbb{E}_{Z\sim P}\left[l(Z;\theta)\right]$$

- batch learning: minimize the empirical loss

$$\hat{\theta}_t = \arg\min_{\theta} \hat{L}_t(\theta),$$

- on-line learning: update sequentially with a datum sampled at each time (or resampled from pooled data)

$$\theta_t = \theta_{t-1} - \Phi_t \nabla l(z_t; \theta_{t-1}),$$

where ∇ denotes the gradient with respect to θ, and Φ is a matrix which controls the rate of convergence.

- batch learning:

  pros can adopt wide class of loss functions

  cons shows slow convergence

  may have many local minima

  should store all the observations

- on-line learning:

  pros do not have to store all the observations

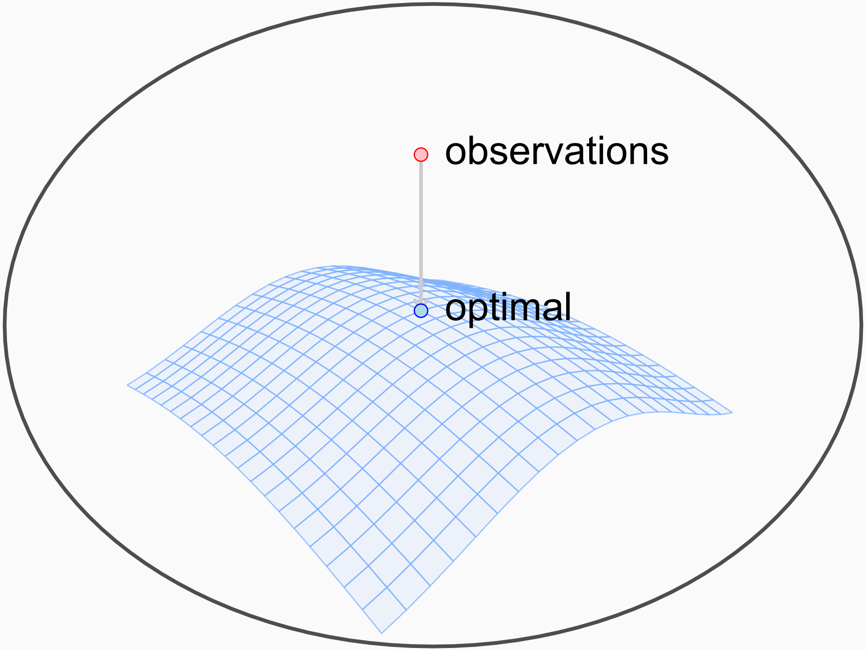
  (good for massive data stream)

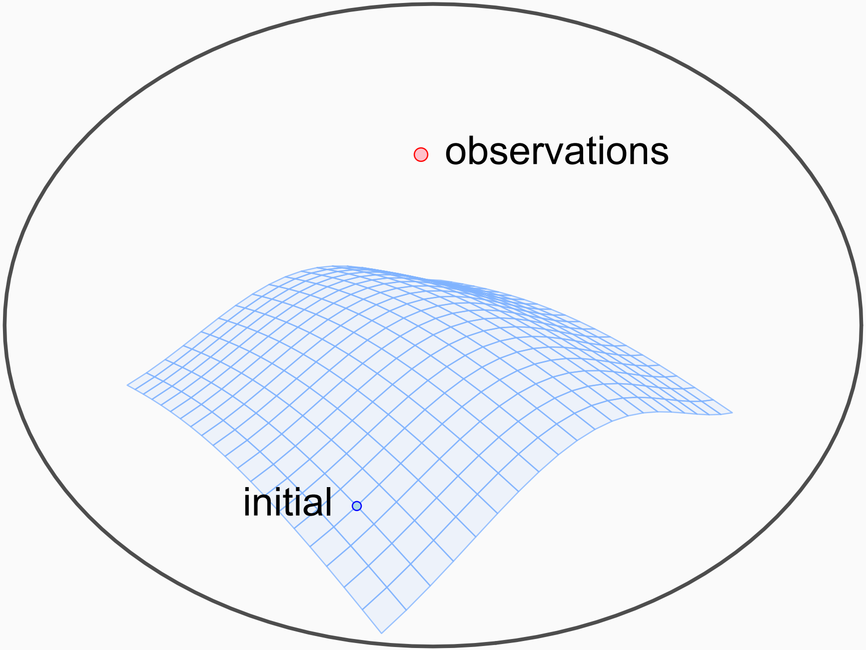
  can escape from local minima

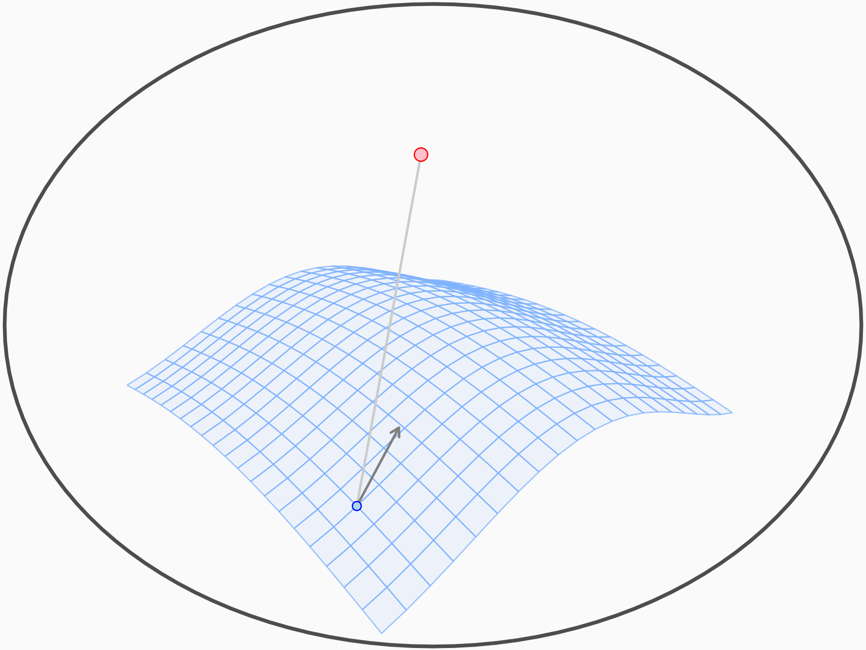
  can follow the change of true distributions

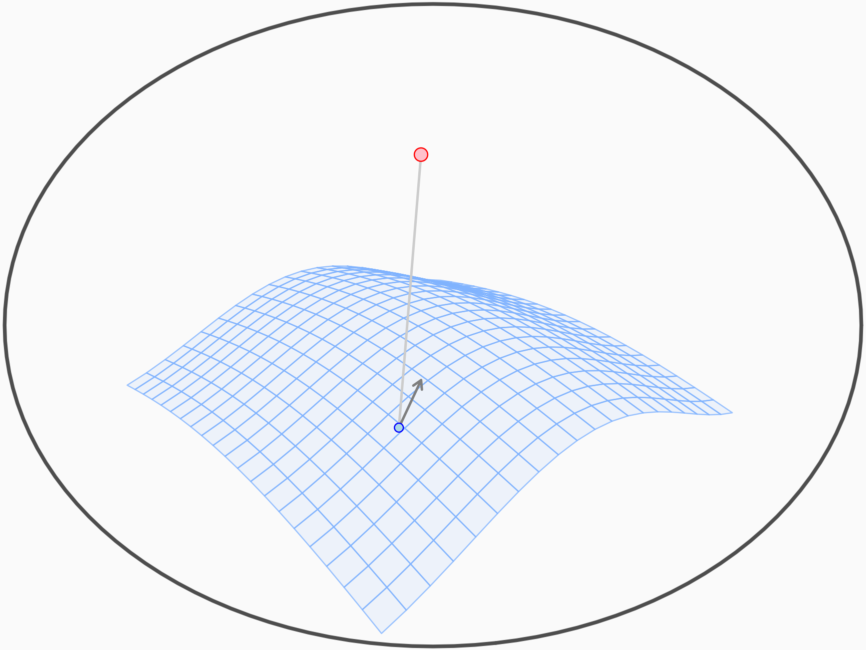
  cons should control learning rate ε properly

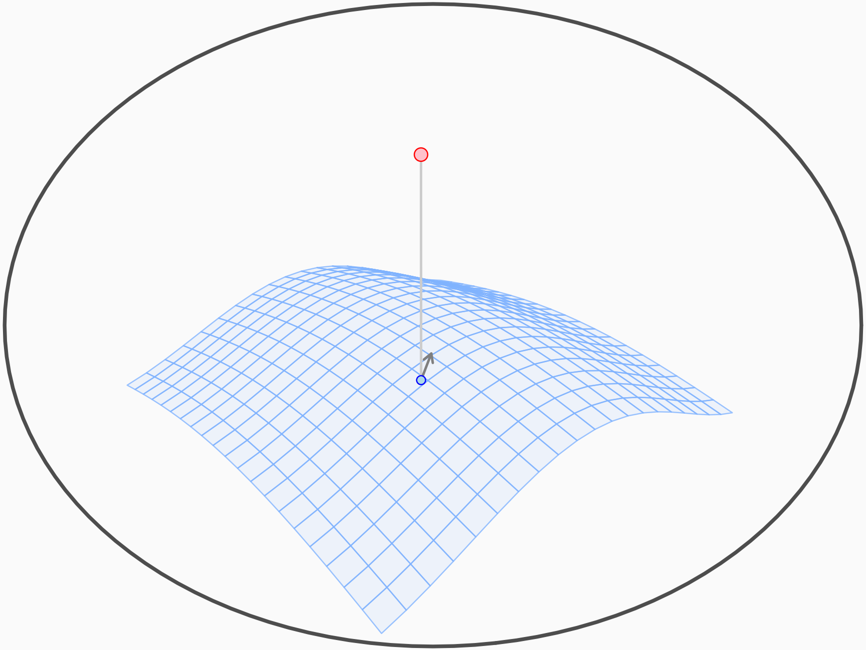
  (do not converge with constant ε)

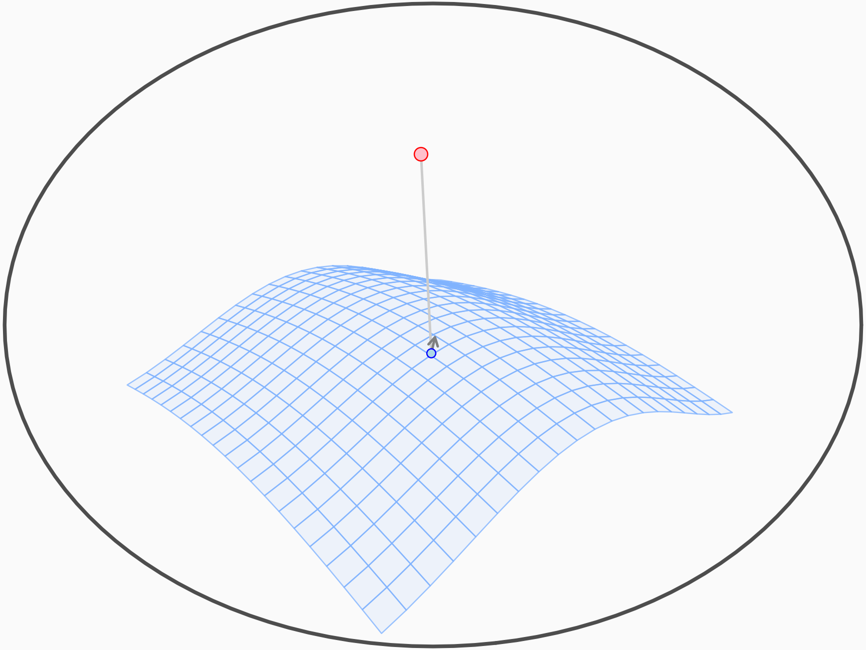

observations
optimal

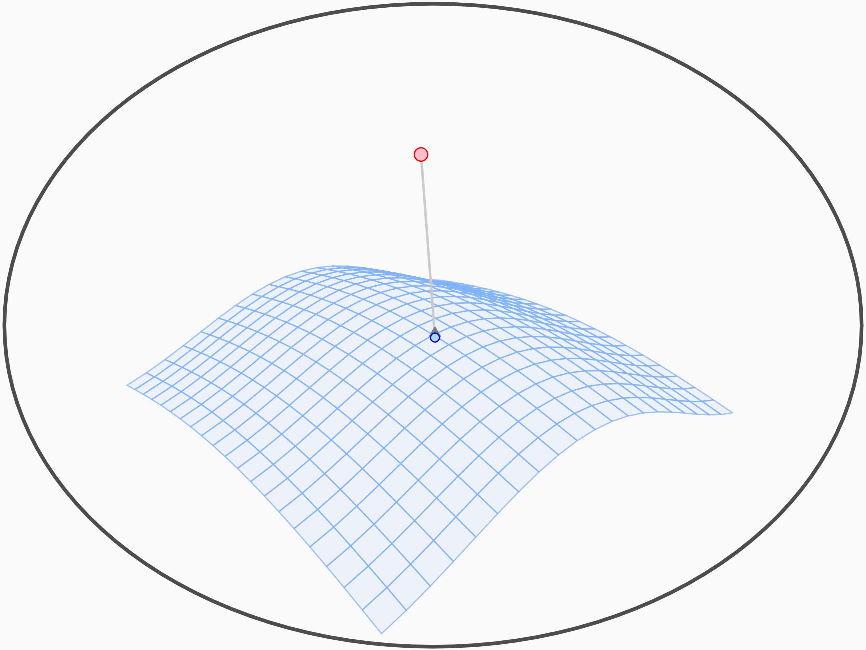

observations
initial

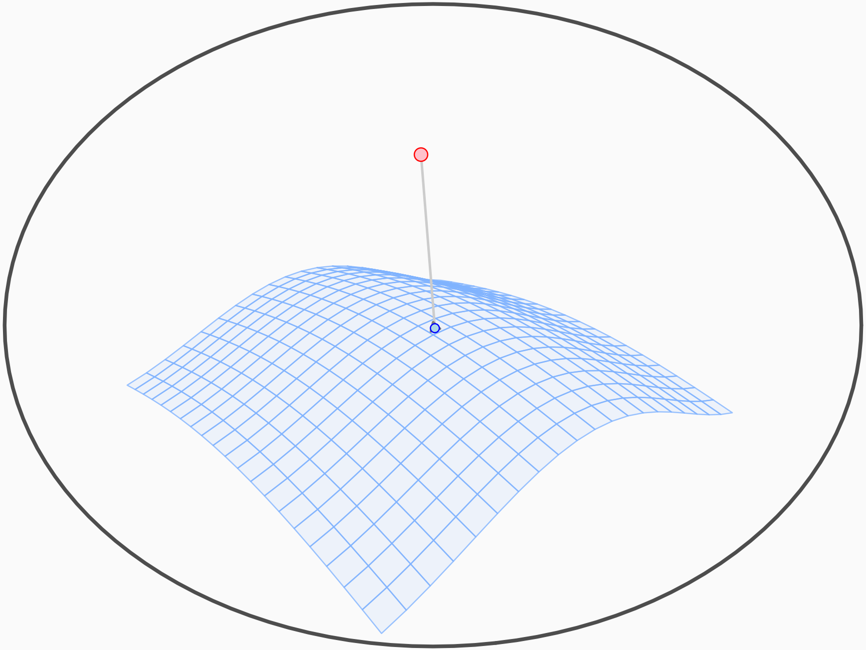


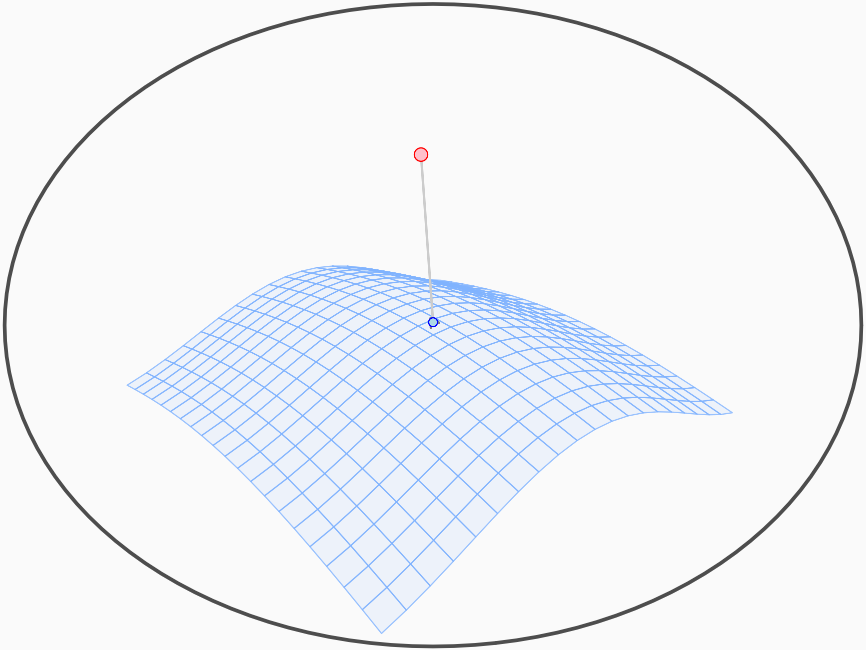


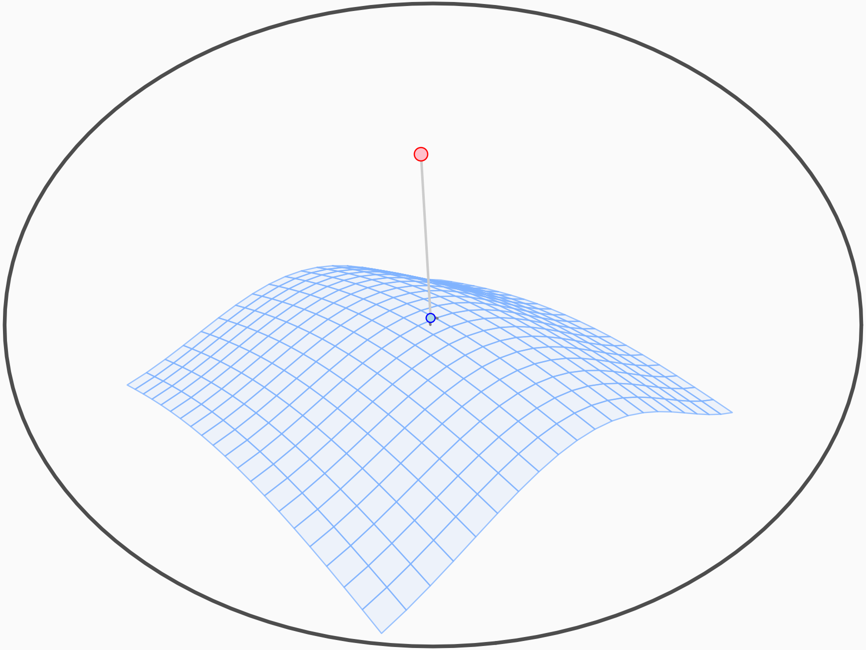


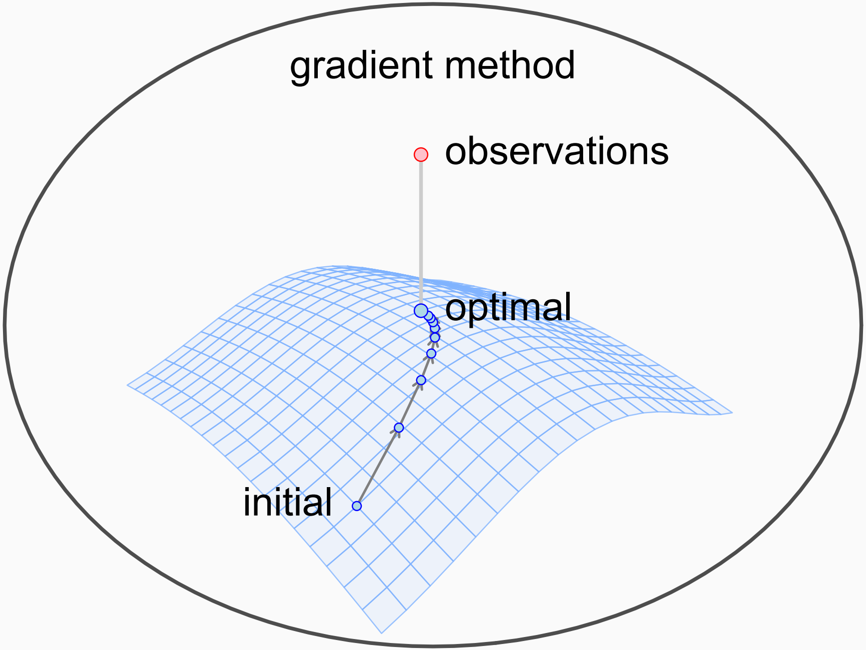


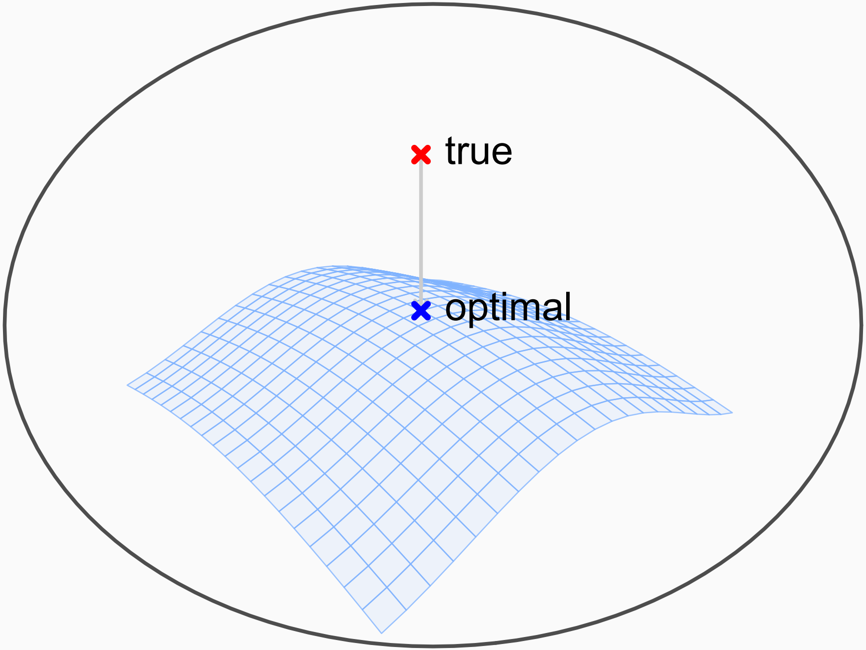


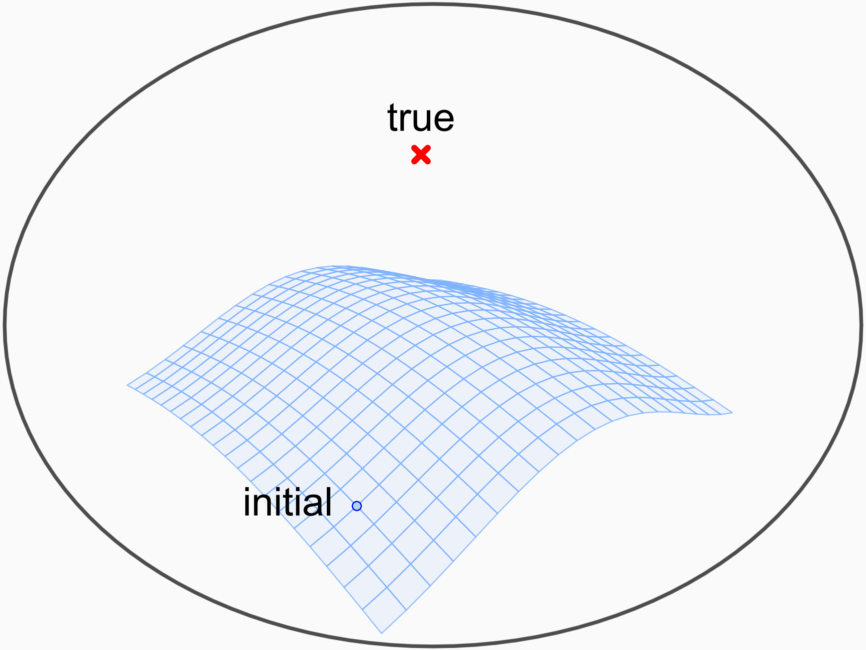


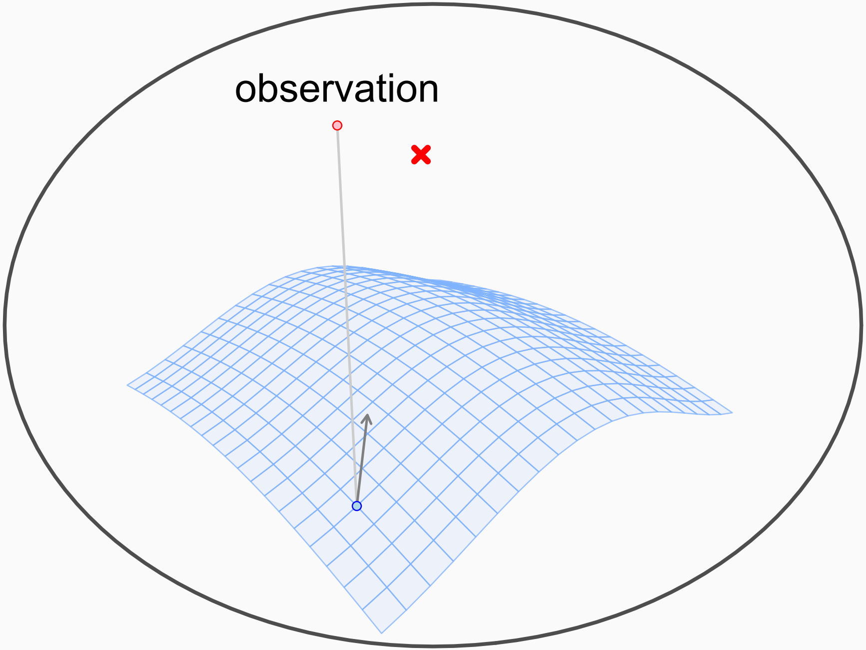


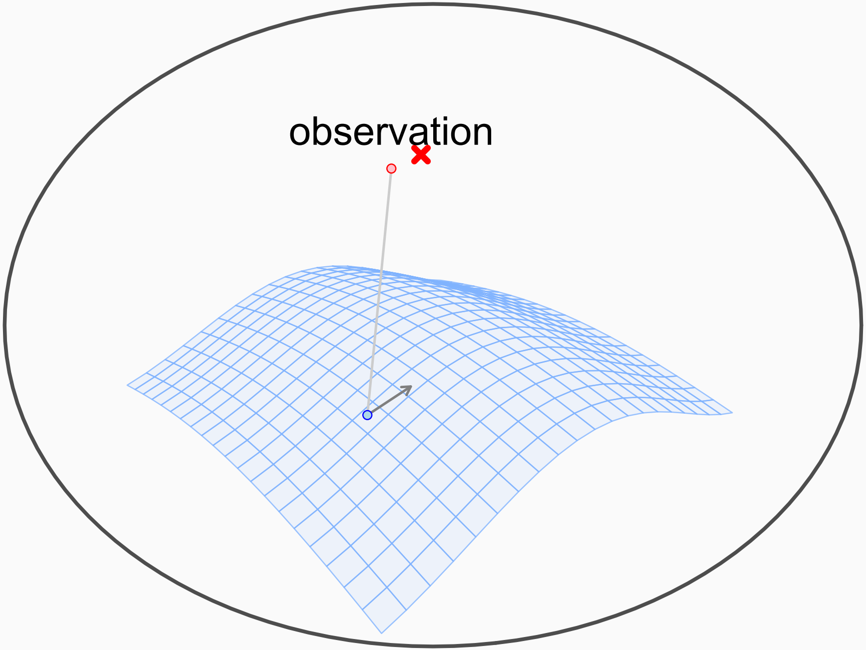


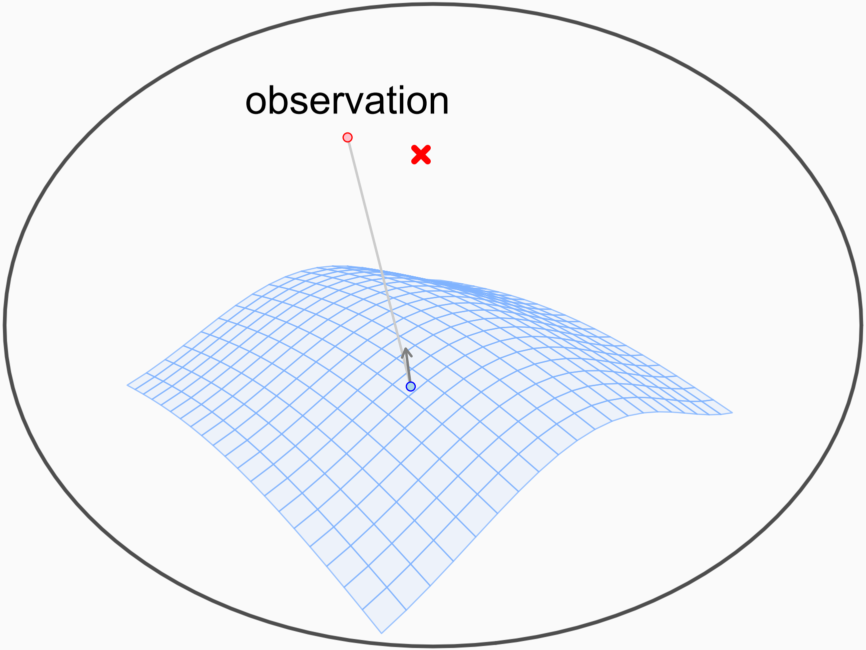

gradient method
observations
optimal
initial

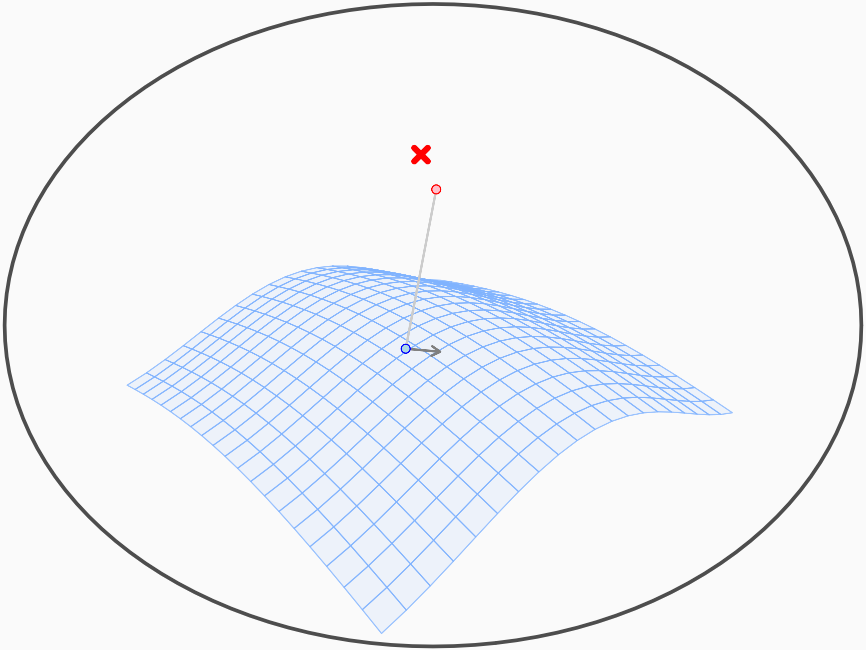

true
optimal

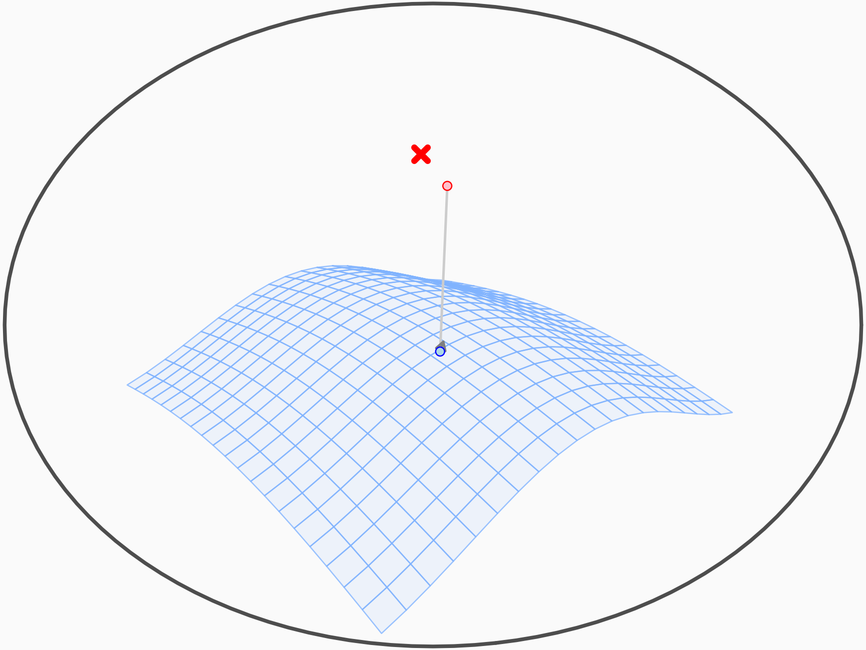

true
initial

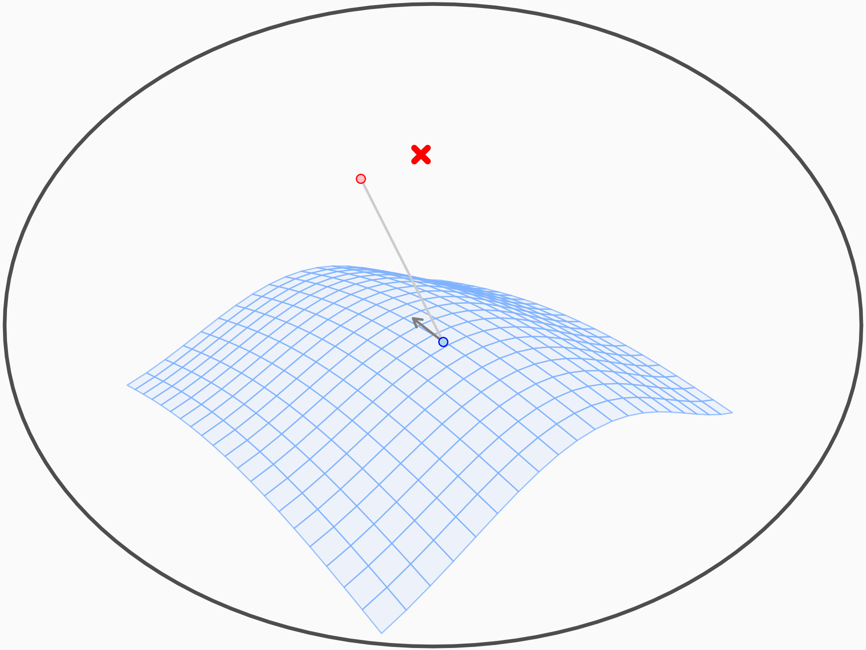

observation

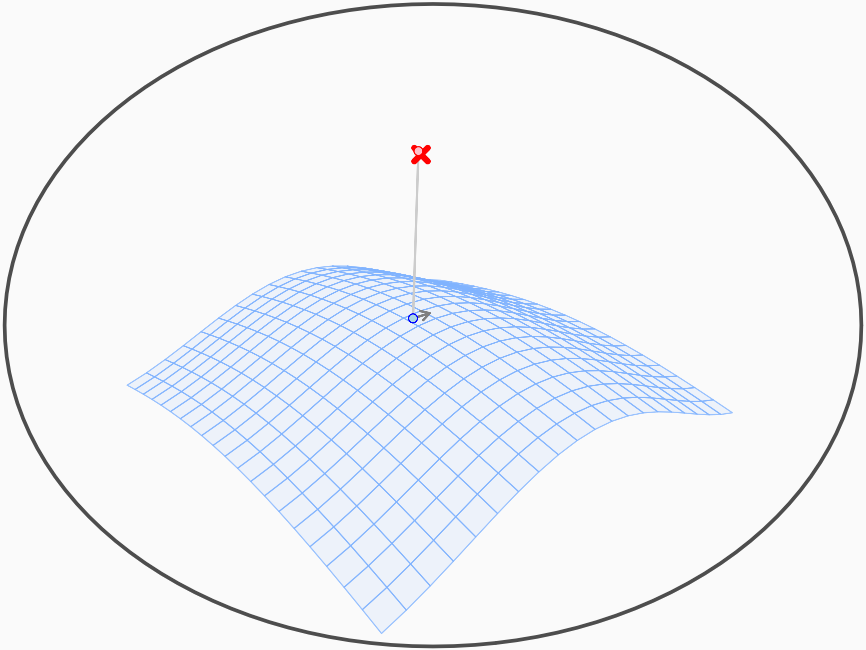

observation

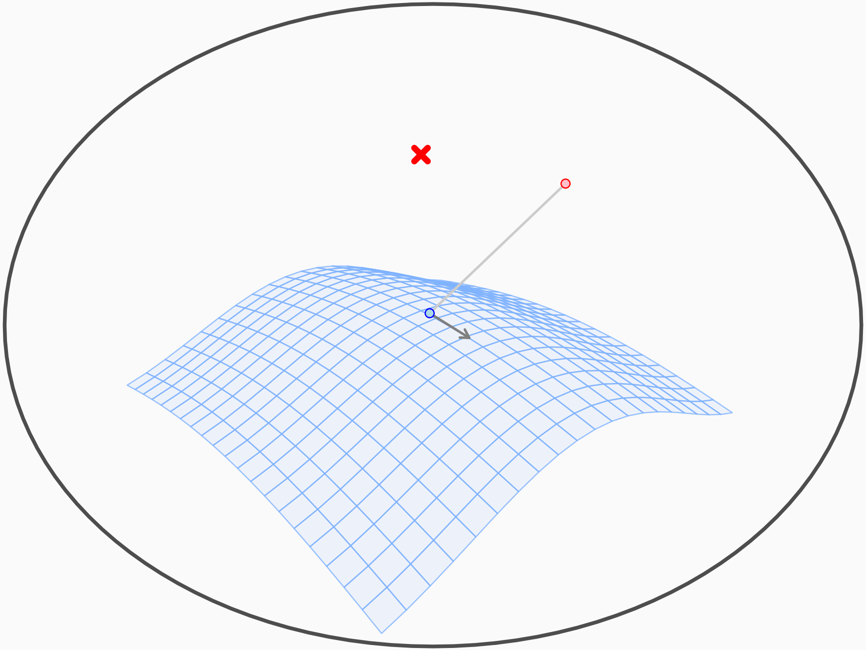

observation

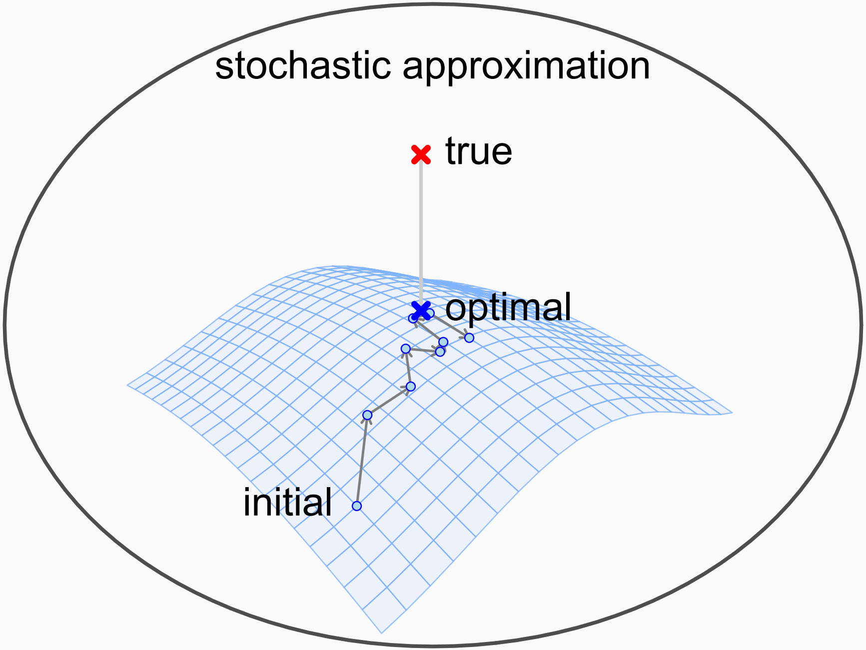











stochastic approximation
true
optimal
initial

- is on-line learning inferior to batch?
- how on-line estimators behave?
- what are good learning parameters?

Problem Formulation

Lemma (Godambe, 1991)

The distribution of $\hat{\theta}_t$ converges to the normal distribution

$$\hat{\theta}_t \sim \mathcal{N}\left(\theta_*, \frac{1}{t}V\right), \quad V = H^{-1}GH^{-1}$$

under some regularity condition, where

$$G = \mathbb{E}_{Z\sim P}\left[\nabla l(Z;\theta)\nabla l(Z;\theta)^\top\right],$$
$$H = \mathbb{E}_{Z\sim P}\left[\nabla\nabla l(Z;\theta)\right],$$

and θ is the optimal parameter of the population loss:

$$\theta = \arg\min_{\theta} L(\theta).$$

Theorem

The expectation of the population loss is asymptotically given by

$$\mathbb{E}\left[L(\hat{\theta}_t)\right] = L(\theta_*) + \frac{1}{2t}\operatorname{tr} GH^{-1} + o\left(\frac{1}{t}\right),$$

where the expectation is taken with respect to D_t.

The variance is asymptotically given by

$$\mathbb{V}\left[L(\hat{\theta}_t)\right] = \frac{1}{2t^2}\operatorname{tr} GH^{-1}GH^{-1} + o\left(\frac{1}{t^2}\right).$$

Theorem

The expectation of the empirical loss is asymptotically given by

$$\mathbb{E}\left[\hat{L}_t(\hat{\theta}_t)\right] = L(\theta) - \frac{1}{2t}\operatorname{tr} GH^{-1} + o\left(\frac{1}{t}\right),$$

where the expectation is taken with respect to D_t.

The variance is asymptotically given by

$$\mathbb{V}\left[\hat{L}_t(\hat{\theta}_t)\right] = \frac{1}{t}\mathbb{V}_{Z\sim P}\left[l(Z;\theta)\right] + o\left(\frac{1}{t}\right).$$

- generalization error:

$$\mathbb{E}\left[L(\hat{\theta}_t)\right] = L(\theta_*) + \frac{1}{2t}\operatorname{tr} GH^{-1} + o\left(\frac{1}{t}\right),$$

- training error:

$$\mathbb{E}\left[\hat{L}_t(\hat{\theta}_t)\right] = L(\theta) - \frac{1}{2t}\operatorname{tr} GH^{-1} + o\left(\frac{1}{t}\right),$$

Corollary (Akaike, 1974)

The generalization error is estimated from the training error by correcting the bias as

$$L(\hat{\theta}_t) = \hat{L}_t(\hat{\theta}_t) + \frac{1}{t}\operatorname{tr} GH^{-1}.$$

In the case of the maximum likelihood estimation, if the ground truth is realized by θ,

$$L(\hat{\theta}_t) = \hat{L}_t(\hat{\theta}_t) + \frac{m}{t} \quad (m: \text{ dim. of } \theta),$$

because $H = G$.

Lemma (Akahira & Takeuchi, 1981; Bottou & Le Cun, 2005)

Let $\hat{\theta}_{t-1}$ and $\hat{\theta}_t$ be estimates for D_{t-1} and $D_t = D_{t-1} \cup \{z_t\}$. Then

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{1}{t}\hat{H}_t^{-1}\nabla l(z_t; \hat{\theta}_{t-1}) + \mathcal{O}_p\left(\frac{1}{t^2}\right)$$

holds under some mild condition, where $\hat{H}_t$ is the empirical Hessian defined by

$$\hat{H}_t = \frac{1}{t}\sum_{z_i \in D_t} \nabla\nabla l(z_i; \hat{\theta}_{t-1}).$$

- batch learning:

$$\hat{\theta}_t = \hat{\theta}_{t-1} - \frac{1}{t}\hat{H}_t^{-1}\nabla l(z_t; \hat{\theta}_{t-1}) + \text{(higher order term)}$$

- optimal on-line learning:

$$\theta_t = \theta_{t-1} - \frac{1}{t}\tilde{H}_{t-1}^{-1}\nabla l(z_t; \theta_{t-1}) + \text{(higher order term)}$$

- optimal design: Newton-Raphson $+ 1/t$-annealing

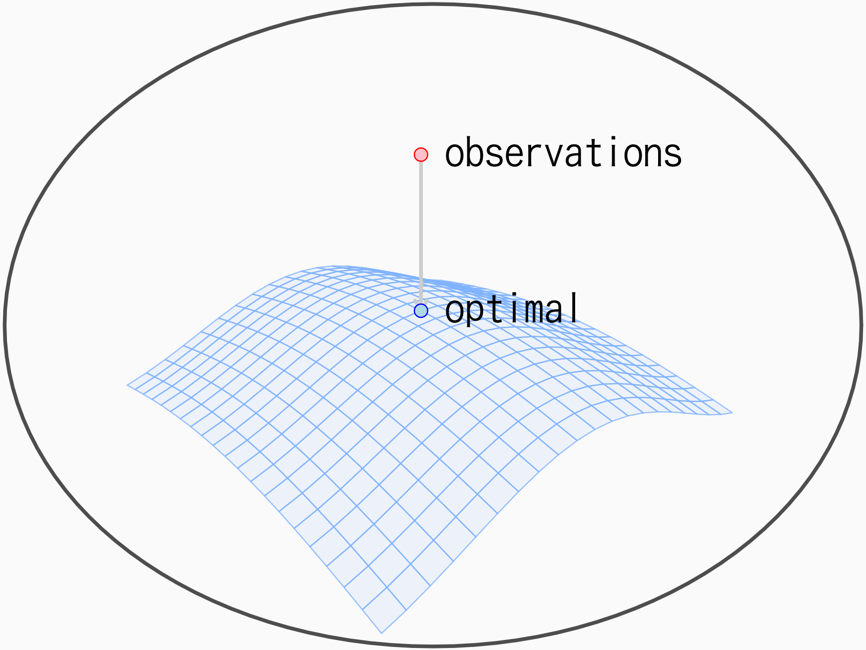
$$\Phi_t = \frac{1}{t}\hat{H}_t^{-1},$$

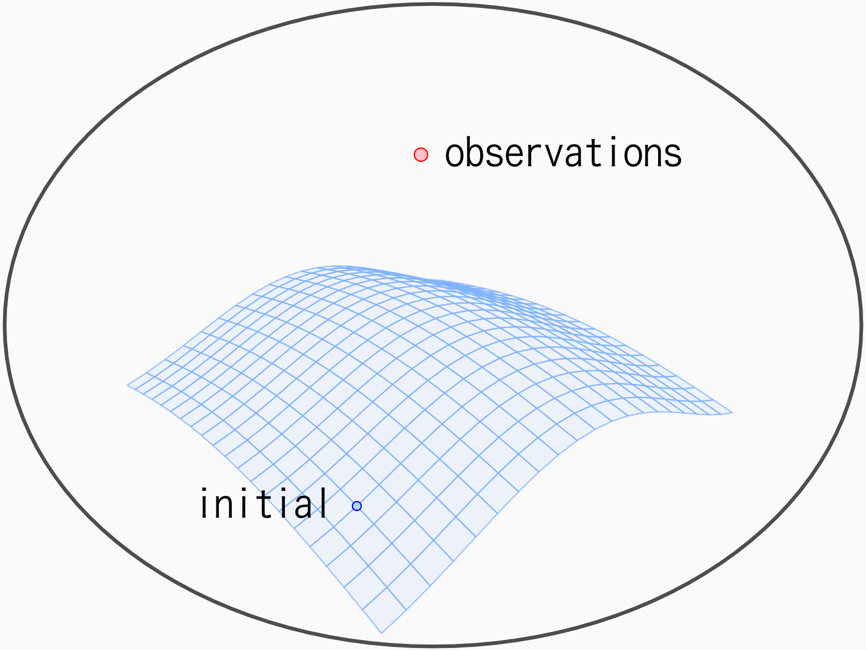
- on-line estimate of Hessian: %(Kalman filtering;Bottou, 1998) (MLE case; Bottou, 1998)

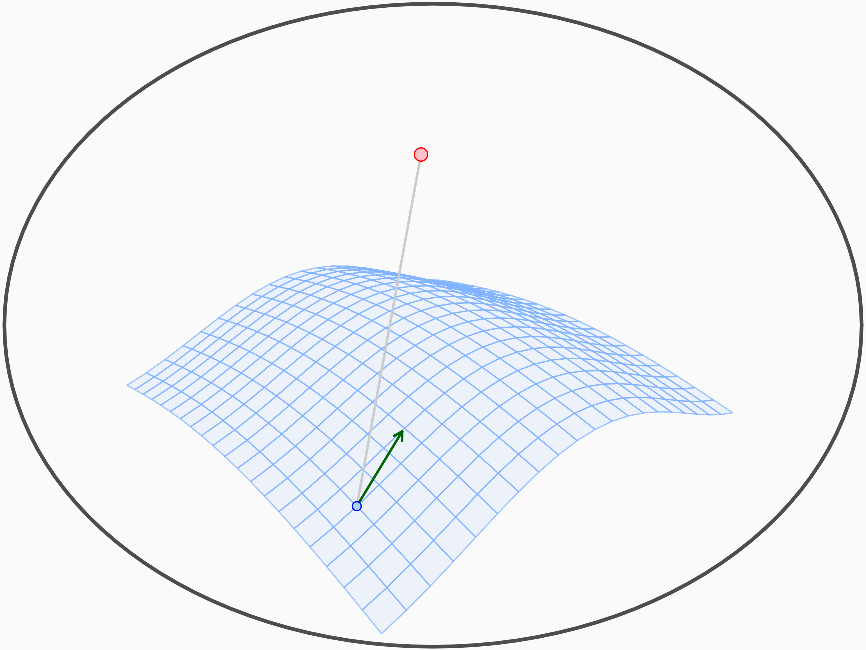
$$\Phi_{t+1} = \Phi_t - \frac{\Phi_t \nabla l \nabla l^\top \Phi_t}{1 + \nabla l^\top \Phi_t \nabla l}$$
$$\text{where } \nabla l = \nabla l(z_{t+1}; \theta_t)$$

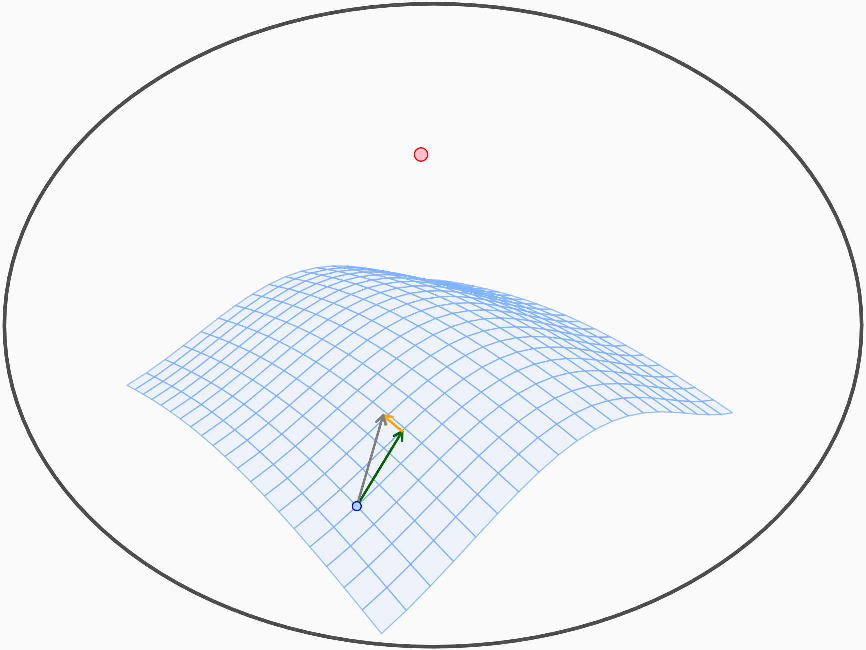
stochastic-BFGS (Nocedal et al, 2014), etc.

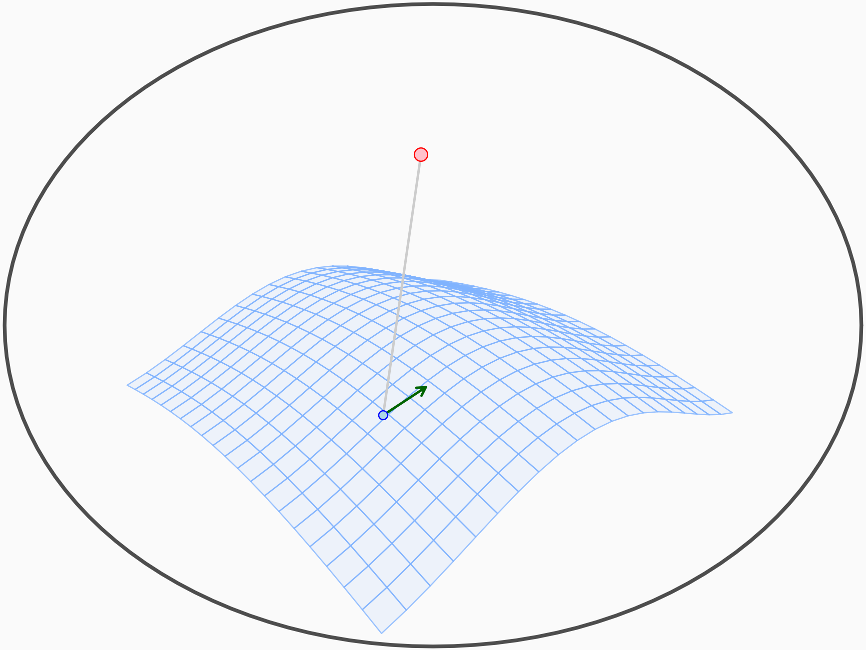
- rate of convergence: equivalent with batch learning
(NM, 1998; NM & Amari, 1999; Bottou & Le Cun, 2005)

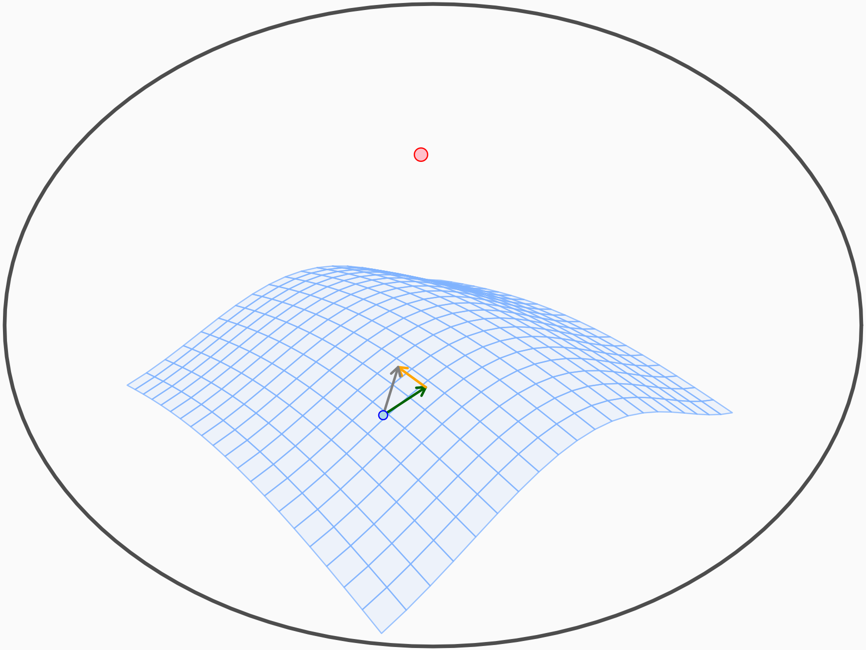

observations
optimal

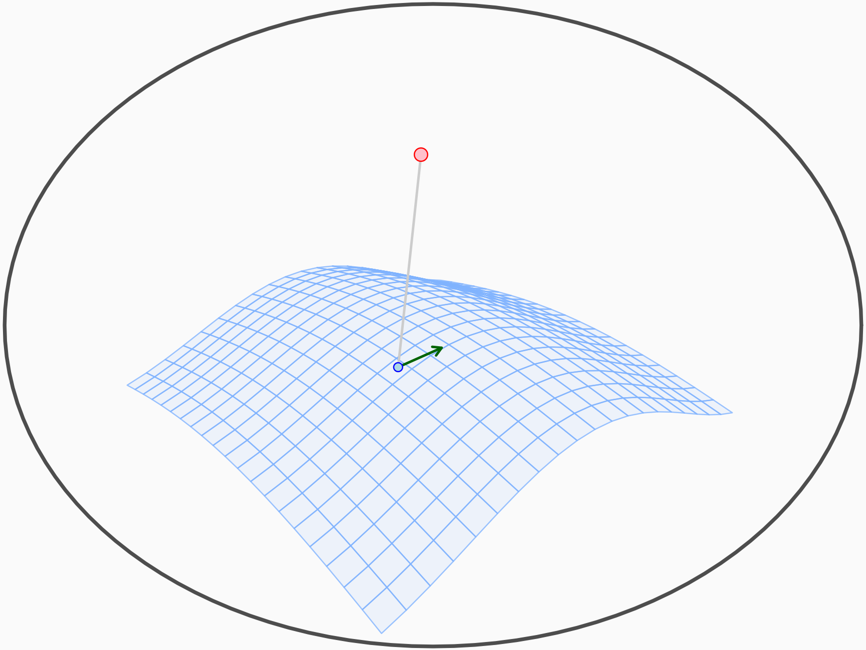

observations
initial

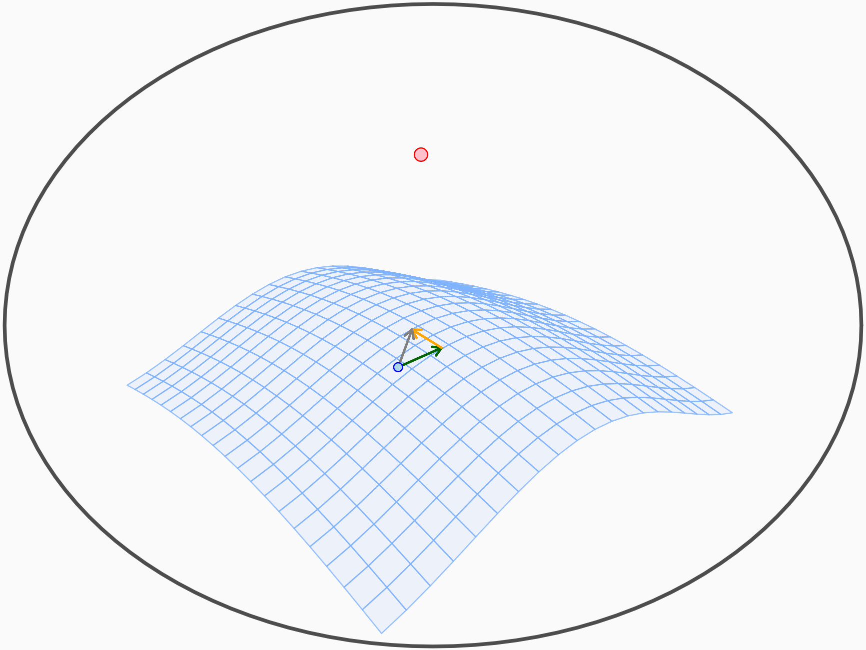


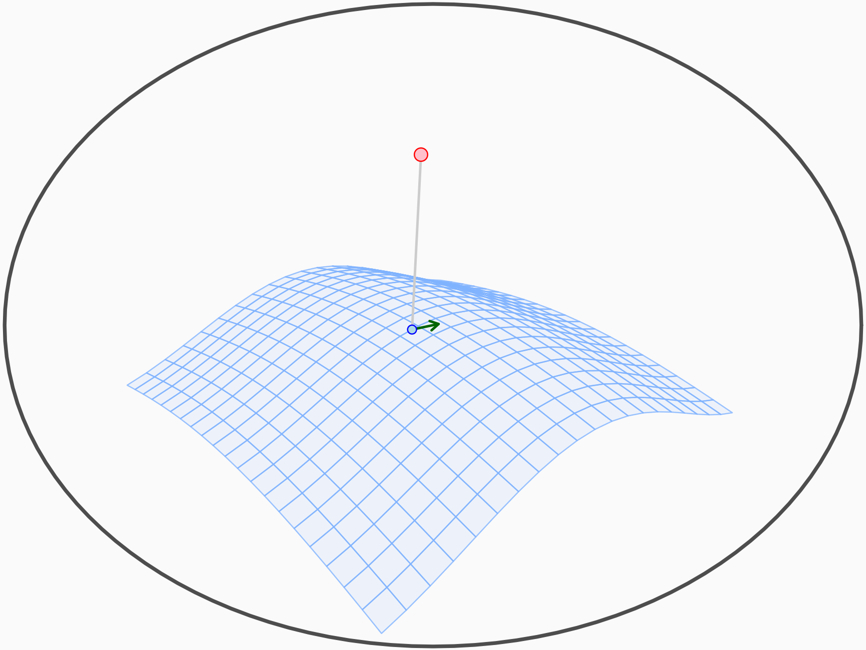


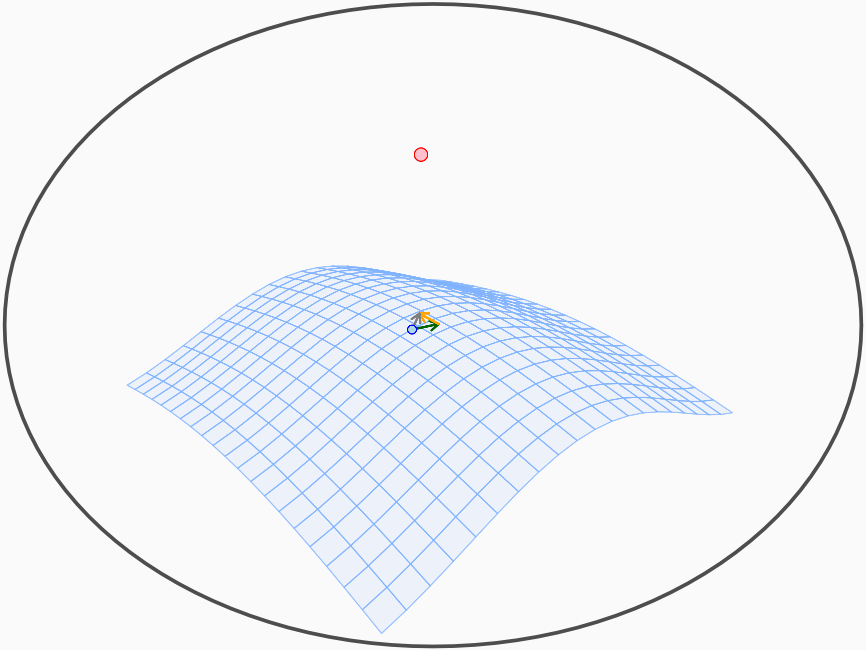


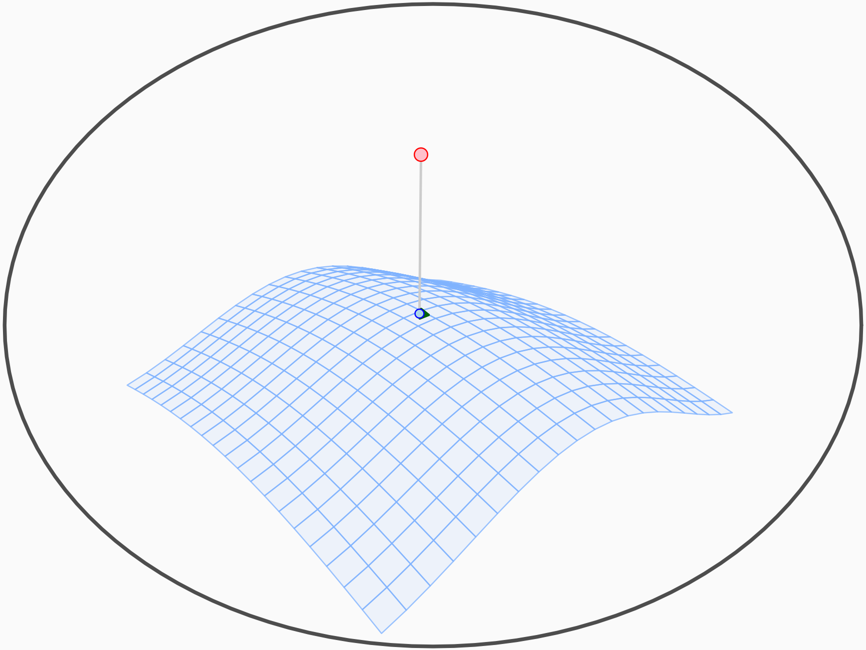


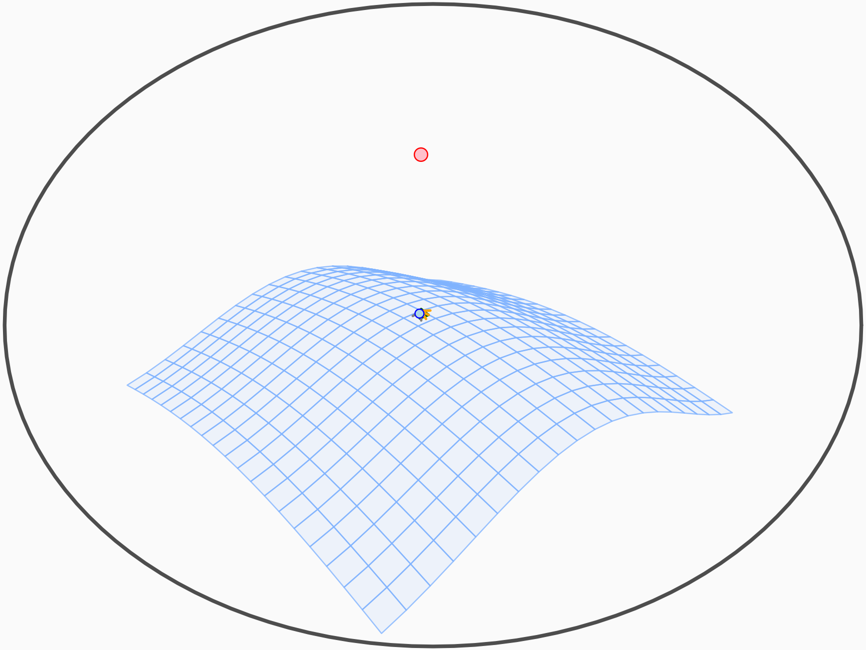


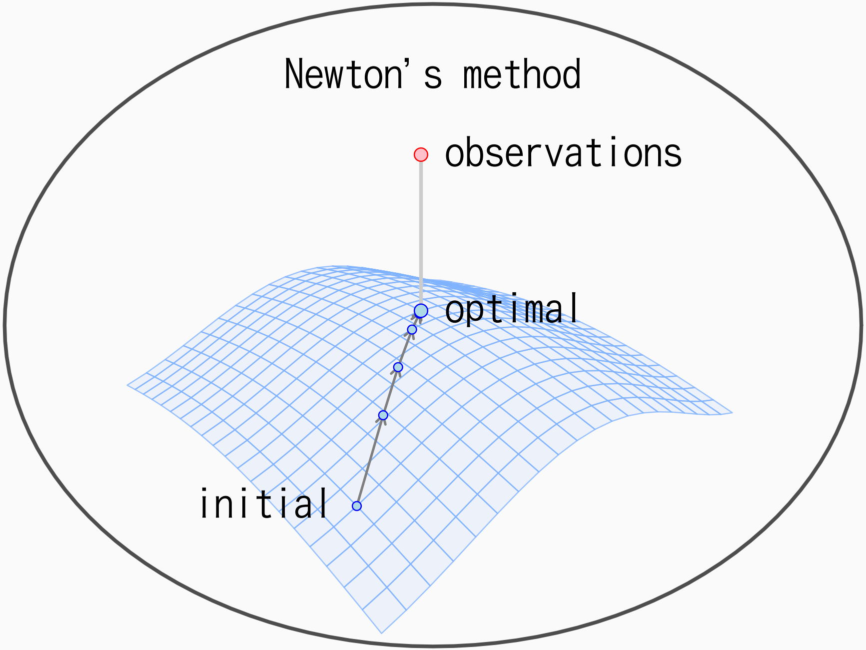











Newton's method
observations
optimal
initial

Lemma (Amari, 1967)

$$\mathbb{E}^{\theta_{t+1}}\left[f(\theta_{t+1})\right] = \mathbb{E}^{\theta_t}\left[f(\theta_t)\right] - \mathbb{E}^{\theta_t}\left[\nabla f(\theta_t)^\top \Phi_t \nabla L(\theta_t)\right] + \frac{1}{2}\operatorname{tr}\mathbb{E}^{\theta_t}\left[\Phi_t G(\theta_t)\Phi_t^\top \nabla\nabla f(\theta_t)\right] + \mathcal{O}(\|\Phi_t\|^3)$$

holds for any smooth function $f(\theta)$, where $\mathbb{E}^{\theta}$ denotes the expectation with respect to θ, and $G(\theta)$ is defined by

$$G(\theta) = \mathbb{E}_{Z\sim P}\left[\nabla l(Z;\theta)\nabla l(Z;\theta)^\top\right].$$

Definition

Let A be an $m \times m$ square matrix and M be an $m \times m$ symmetric matrix. We define two linear operators as follows:

$$\Xi_A M = AM + (AM)^\mathsf{T},$$
$$\Omega_A M = AMA^\mathsf{T}.$$

Lemma

Around the optimal parameter, the following approximated recursive relations for the expectation $\bar{\theta}_t = \mathbb{E}^{\theta_t}[\theta_t]$ and the covariance $V_t = \mathbb{V}^{\theta_t}[\theta_t]$ hold:

$$\begin{aligned}\bar{\theta}_{t+1} &= \bar{\theta}_t - Q_t(\bar{\theta}_t - \theta),\\ V_{t+1} &= V_t - \Xi_{Q_t} V_t + \Omega_{Q_t} V - \Omega_{Q_t}(\bar{\theta}_t - \theta)(\bar{\theta}_t - \theta)^\top,\end{aligned}$$

where

$$Q_t = \Phi_t H, \quad V = H^{-1} G H^{-1}.$$

(note: $\Xi_A M = AM + (AM)^\top$, $\Omega_A M = AMA^\top$)

Theorem

Let Φ be C/t, where C is a constant matrix. If $\lambda_{\min}(CH) \geq 1$, the leading terms are given by

$$\bar{\theta}_t = \theta + S_t(\theta_0 - \theta), \quad S_t = \prod_{\tau=2}^{t} \left(I - \frac{CH}{\tau} \right) = \mathcal{O}\left(\frac{1}{t^{\lambda_{\min}}} \right),$$

$$V_t = \left[\left(\Xi_{CH} - I \right)^{-1} \Omega_{CH} \right] \frac{1}{t} V, \quad V = H^{-1} G H^{-1},$$

where θ_0 is an initial parameter.

Lemma

Let $\lambda_i,\ i = 1, \ldots, m$ be eigenvalues of A. The eigenvalues of Ξ_A and Ω_A are given by

$$\Xi_A : \lambda_i + \lambda_j,\ i, j = 1, \ldots, m,$$
$$\Omega_A : \lambda_i \lambda_j,\ i, j = 1, \ldots, m.$$

.

This follows by the relation

$$\mathrm{cs}(ABC) = (C^\top \otimes A)\mathrm{cs}B$$

for any $m \times m$ square matrices A, B, C. □

- larger $\lambda_{\min}$ is advantageous to faster convergence of $\bar{\theta}_t$.
- $(\Xi_{CH} - I)^{-1}\Omega_{CH}$ expands V/t, which is the minimum covariance attained by batch learning.
- eigenvalues of $(\Xi_{CH} - I)^{-1}\Omega_{CH}$ are given by

$$\frac{\lambda_i \lambda_j}{\lambda_i + \lambda_j - 1},$$

where λ_i's are eigenvalues of CH.

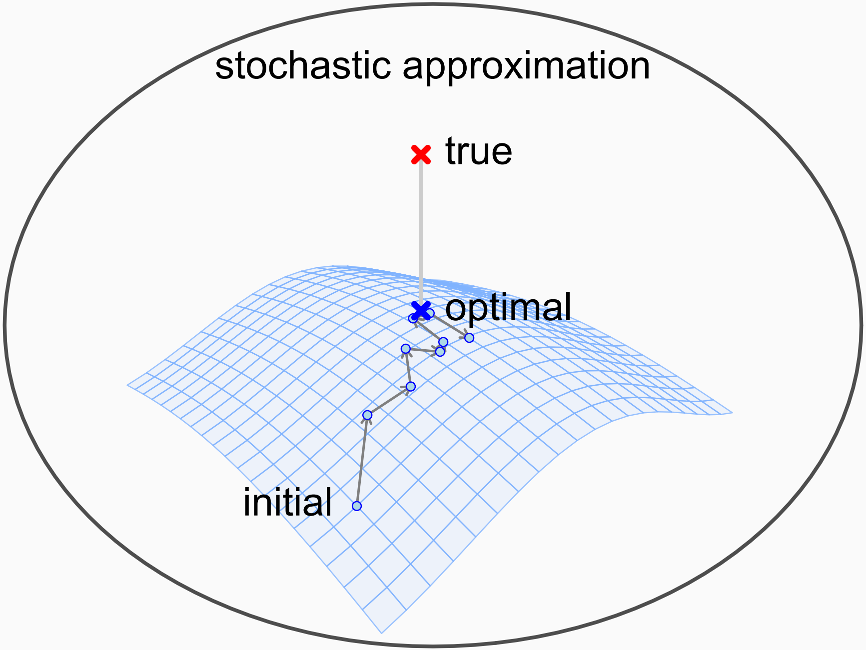
- if $C = H^{-1}$, %i.e. $CH = I$, all the eigenvalues of $(\Xi_I - I)^{-1}\Omega_I$ are equal to 1, i.e. $V_t = V/t$.
- $\Phi_t = H^{-1}/t$ is optimal.

- on-line learning:

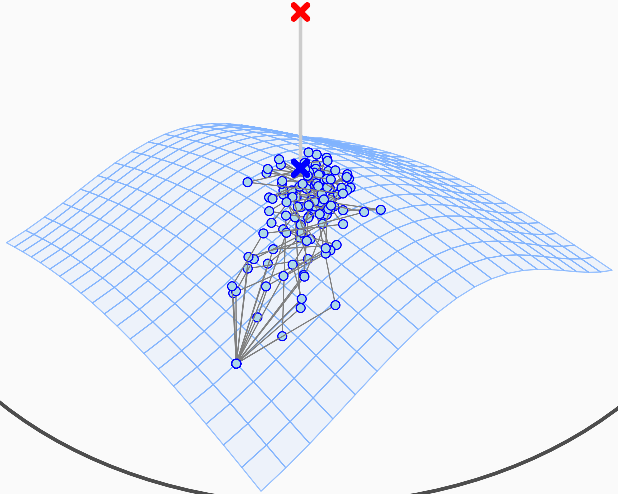
$$\mathbb{E}\left[(\theta_t - \theta)(\theta_t - \theta)^\top\right] = \mathbb{V}\left[\theta_t\right] + \mathbb{E}\left[\theta_t - \theta\right]\mathbb{E}\left[\theta_t - \theta\right]^\top$$
$$= \frac{1}{t}V + \mathcal{O}\left(\frac{1}{t^2}\right).$$

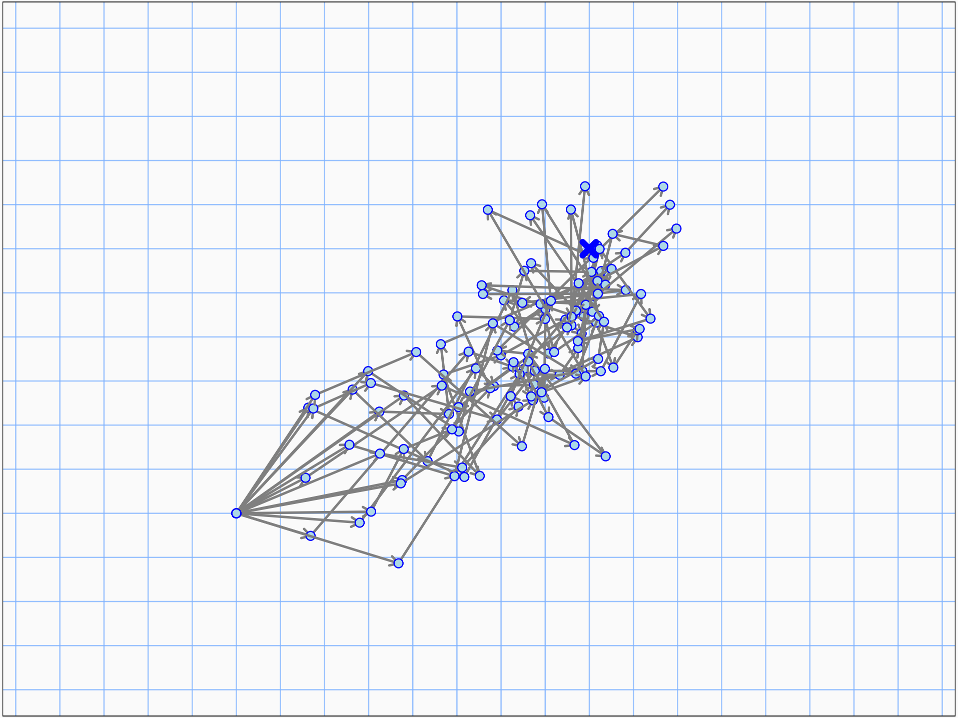
- batch learning:

$$\mathbb{E}\left[(\hat{\theta}_t - \theta)(\hat{\theta}_t - \theta)^\top\right] = \frac{1}{t}V + \mathcal{O}\left(\frac{1}{t^2}\right).$$

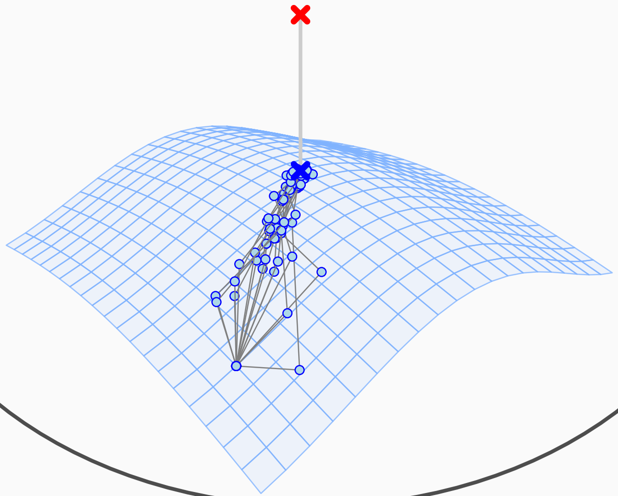

stochastic approximation
true
optimal
initial

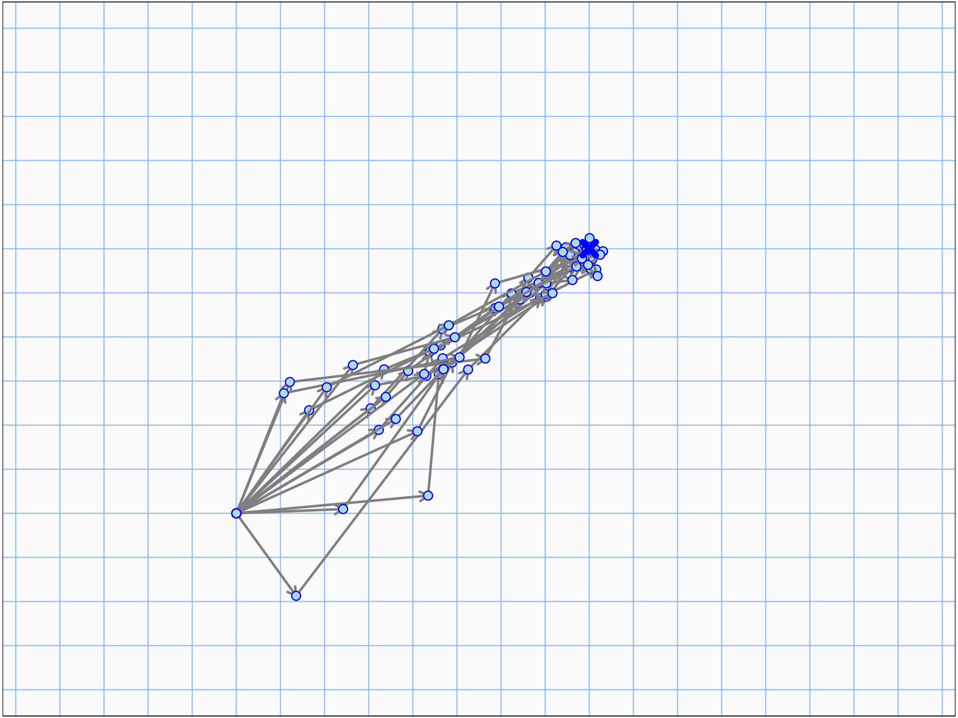
fixed learning rate





optimal learning rate





Illustrative Example

a method for evaluating the relative skill levels of players

- Elo rating: Arpad Elo, 1960
  used in competitor-versus-competitor games such as chess
  scores given to players are updated according to game results
- Glicko rating: Mark Glickman, 1997
  including confidence of estimated skill levels
- TrueSkill: Ralf Herbrich et al., 2007
  extension to multiplayer games
  skill levels are random variables (Bayesian framework)

- score: $\theta = (\theta^1, \theta^2, \dots)$
- event: $z_t = (a \succ b)$ (player a beats player b at time t)
- probability model:

$$\Pr(a \succ b) = P(z_t; \theta) = \frac{1}{1 + \exp(\gamma \cdot (\theta^b - \theta^a))},$$

  where γ is defined such that a player whose rating is 200 points greater than the other is expected to have a 75\
- loss function: (negative log loss)

$$l(z_t; \theta) = -\log P(z_t; \theta) = \log(1 + \exp(\gamma \cdot (\theta^b - \theta^a)))$$

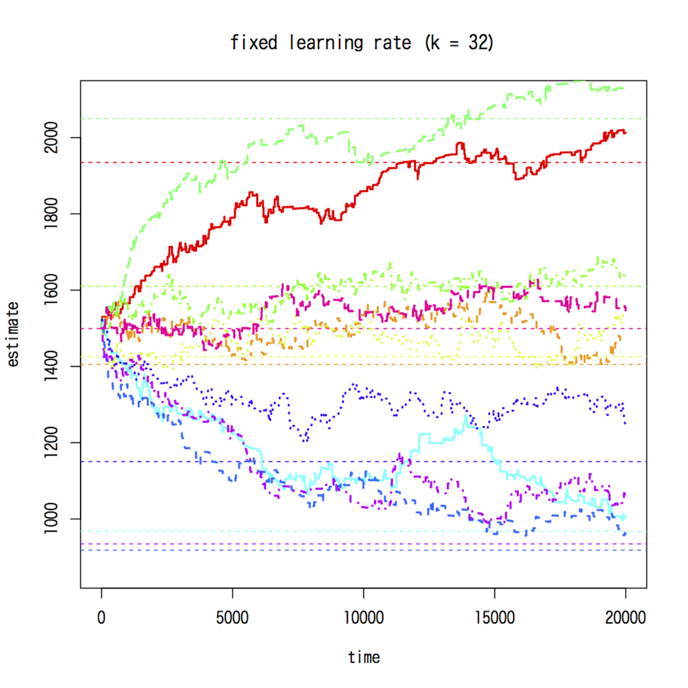
- gradient:

$$\frac{\partial}{\partial \theta^i} l(z_t;\theta) = \begin{cases} 0, & i \neq a, b \\ -\gamma \cdot (1 - P(z_t;\theta)), & i = a \text{ (winner)} \\ +\gamma \cdot (1 - P(z_t;\theta)), & i = b \text{ (looser)} \end{cases}$$

- update rule:

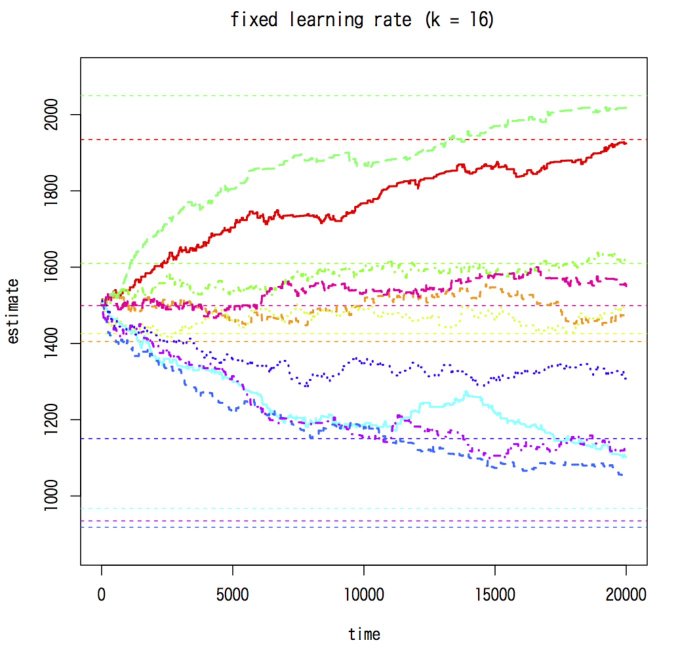
$$\begin{aligned} \theta_{t+1} &= \theta_t - \varepsilon \nabla l(z_t;\theta) \\ &= \theta_t + (0, \ldots, \underbrace{\varepsilon\gamma(1-P)}_{a}, \ldots, \underbrace{-\varepsilon\gamma(1-P)}_{b}, \ldots, 0)^T \end{aligned}$$

where $k = \varepsilon\gamma =$ 32 for novices, 16 for professionals.



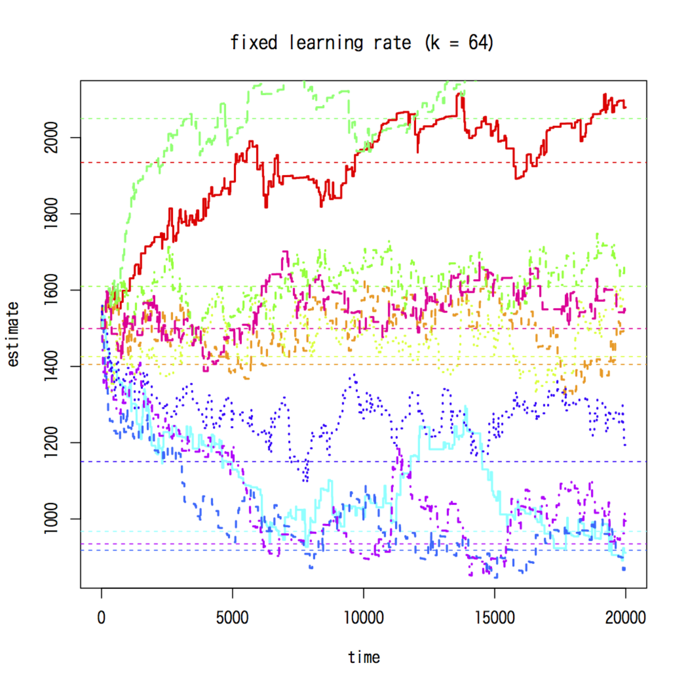
fixed rate\ $\Phi_t = \varepsilon I$

- 10 players out of 100
- 20000 games {(400[games/pl.])}
- $k = 32, 16, 64$
- $\theta_0^i = 1500$



fixed rate\ $\Phi_t = \varepsilon I$

- 10 players out of 100
- 20000 games {(400[games/pl.])}
- $k = 32, 16, 64$
- $\theta_0^i = 1500$



fixed rate\ $\Phi_t = \varepsilon I$

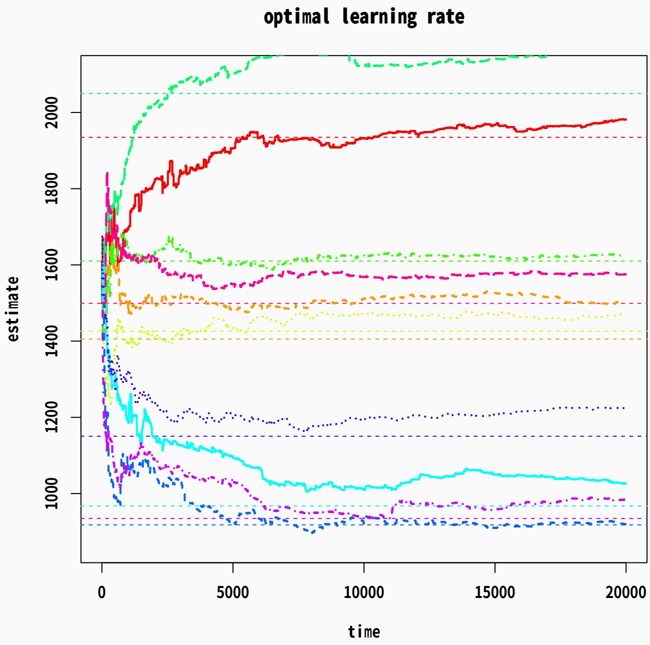
- 10 players out of 100
- 20000 games $\{(400[\text{games/pl.}])\}$
- $k = 32, 16, 64$
- $\theta_0^i = 1500$

- update rule: (Φ: matrix)

$$\begin{aligned}
\theta_{t+1} &= \theta_t - \Phi_t \nabla l(z_t;\theta_t),\\
\Phi_{t+1} &= \Phi_t - \frac{\Phi_t \nabla l_t \nabla l_t^\top \Phi_t}{1 + \nabla l_t^\top \Phi_t \nabla l_t},\\
\nabla l_t &= \nabla l(z_{t+1};\theta_t)\\
&= (0,\ldots,\underbrace{\gamma(1-P)}_{a},\ldots,\underbrace{-\gamma(1-P)}_{b},\ldots,0)^T
\end{aligned}$$

- initial value:

$$\Phi_0 = kI \quad I \text{ is the identity matrix}$$



optimal rate

- 10 players out of 100
- 20000 games {(400[games/pl.])}
- sensitive to initial value

- original update rule: $\Delta\theta = -\varepsilon \nabla l(z_t; \theta)$
  - only related players are updated: $\Delta\theta^i = 0, \; i \neq a, b$.
  - sum of θ is kept constant: $\mathbf{1}^\top \Delta\theta = 0$.
- optimal update rule: $\Delta\theta = -\Phi_t \nabla l(z_t; \theta)$
  - all the players are updated, because $\Phi_t = \hat{H}_t^{-1}/t$ is a dense matrix.
  - sum of θ is not necessarily kept constant.
- our problem: design Φ_t to fit the original restriction.

- 1 vs 1 case: (players a and b)

$$\Delta\theta = \alpha \boldsymbol{a}, \quad \boldsymbol{a}^\top = \begin{array}{c} \begin{matrix} a & b & c & \end{matrix} \\ \begin{pmatrix} 1 & -1 & 0 & \cdots \end{pmatrix} \end{array},$$

or

$$B^\top \Delta\theta = 0, \quad B^\top = \begin{array}{c} \begin{matrix} a & b & c & d & \end{matrix} \\ \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & & & \ddots \end{pmatrix} \end{array}.$$

- 2 vs 2 case: (players a+b and c+d)

$$\Delta\theta = A\alpha, \quad A^\top = \begin{matrix} a & b & c & d & e & \\ \end{matrix}$$
$$\begin{pmatrix} 1 & 0 & -1 & 0 & 0 & \cdots \\ 1 & 0 & 0 & -1 & 0 & \cdots \\ 0 & 1 & -1 & 0 & 0 & \cdots \end{pmatrix},$$

or

$$B^\top\Delta\theta = 0, \quad B^\top = \begin{matrix} a & b & c & d & e & f & \\ \end{matrix}$$
$$\begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 1 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 1 & \cdots \\ \vdots & \vdots & & & & & \ddots \end{pmatrix}.$$

Problem A

Find an "optimal" gradient $\Delta\theta = \Phi\nabla l(z;\theta)$ subject to

$$\Delta\theta \in \operatorname{Im} A, \quad (\Delta\theta = A\alpha,\ \alpha \in \mathbb{R}^k)$$

for a matrix $A \in \mathbb{R}^{m\times k}$.

Problem B

Find an "optimal" gradient $\Delta\theta = \Phi\nabla l(z;\theta)$ subject to

$$\Delta\theta \in \operatorname{Ker} B^\top, \quad (B^\top\Delta\theta = 0)$$

for a matrix $B \in \mathbb{R}^{m\times(m-k)}$,

cf. $f(\theta) = \text{const.} \Rightarrow \nabla f(\theta)^\top\Delta\theta = 0$

- optimality is defined in terms of

$$\text{minimize } \|H^{-1}\nabla l - \Delta\theta\|_M,$$

where $\|x\|_M^2 = \langle x, x\rangle_M$ and $\langle x, y\rangle_M = \langle Mx, y\rangle$.

- M is chosen as H, because
  - quadratic approximation of population loss:

$$\|\theta - \theta\|_H^2 = (\theta - \theta)^\top H(\theta - \theta) = L(\theta) - L(\theta)$$

  - Mahalanobis distance in maximum likelihood case:

$$\mathbb{V}[\hat{\theta}_t] = \frac{1}{t}H^{-1}GH^{-1} = \frac{1}{t}H^{-1}$$

% - (Φ_t becomes symmetric.)

- decompose Φ_t into scalar and matrix parts as

$$\Phi_t = \varepsilon_t C, \quad (\text{e.g., } \varepsilon_t = 1/t)$$

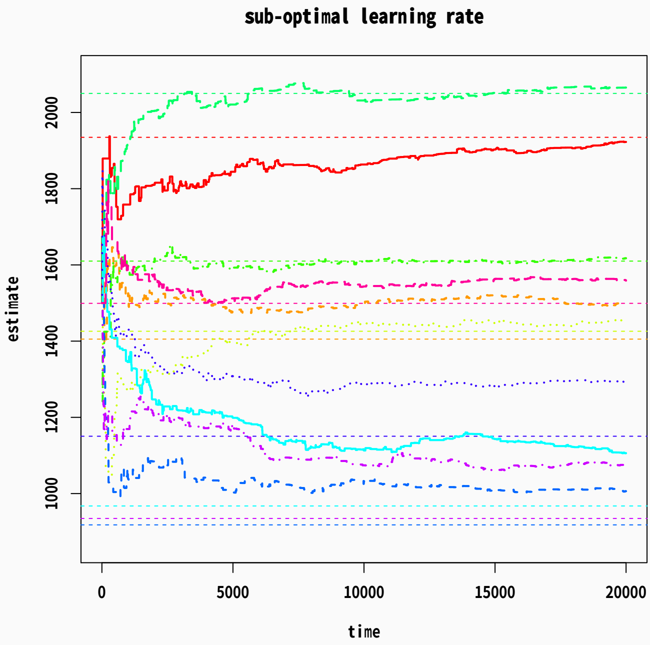
- solutions for the problems are:

Problem A

$$C_A = A(A^\top HA)^{-1}A^\top$$

Problem B

$$C_B = H^{-1} - H^{-1}B(B^\top H^{-1}B)^{-1}B^\top H^{-1}$$



sub-optimal rate

- 10 players out of 100
- 20000 games {(400[games/pl.])}

- C_A and C_B are symmetric (only when $M = H$).
- $C_A H$ or $C_B H$ is a projection matrix:

$$\lambda = \begin{cases} 1, & v \in \mathrm{Im}\, A \text{ or } \mathrm{Ker}\, B, \\ 0, & \text{otherwise.} \end{cases}$$

- if k is small, calculating C_A is more efficient than C_B.
- only a few parameters are updated, however convergence is as good as optimal case.
  (information loss is quite small in some case)

Conclusion

we have investigated

- dynamics of convergence phase of on-line learning,
- conditions for optimal convergence rate,
- optimal projection of gradients to subspaces,

practical applications would be

- skill level rating systems,
- on-line learning for Bradley-Terry model,
- distributed control systems.

Amari, Shun-ichi (June 1967). "A Theory of Adaptive Pattern Classifiers." In: *IEEE Transactions on Electronic Computers* EC-16 (3), pp. 299–307. DOI: 10.1109/PGEC.1967.264666.

Bottou, Léon (1998). "Online Learning and Stochastic Approximations." In: *Online Learning in Neural Networks*. Ed. by David Saad. Cambridge, UK: Cambridge University Press, pp. 9–42. Google Books: iu2v6C5nx4oC.

Bottou, Léon and Yann LeCun (Mar. 23, 2005). "On-line learning for very large data sets." In: *Applied Stochastic Models in Business and Industry* 21 (2), pp. 137–151. DOI: 10.1002/asmb.538.

Godambe, Vidyadhar P., ed. (Aug. 15, 1991). ***Estimating Functions***. Oxford Statistical Science Series.

Murata, Noboru (1998). "A Statistical Study on On-line Learning." In: *Online Learning in Neural Networks*. Ed. by David Saad. Cambridge, UK: Cambridge University Press, pp. 63–92. Google Books: iu2v6C5nx4oC.

Murata, Noboru and Shun-ichi Amari (Apr. 1999). "Statistical analysis of learning dynamics." In: *Signal Processing* 74 (1), pp. 3–28. DOI: 10.1016/S0165-1684(98)00206-0.