Methods of optimization

Lecturer:

Valery Reizlin,

Associate Professor of IAD

TPU

What is Optimization?

Find the minimum or maximum of an objective function given a set of constraints:

$$\arg\min_x f_0(x)$$
$$\text{s.t.} f_i(x) \leq 0, i = \{1, \ldots, k\}$$
$$h_j(x) = 0, j = \{1, \ldots l\}$$

Why Do We Care?

Linear Classification

$$\arg\min_{w} \sum_{i=1}^{n} ||w||^2 + C \sum_{i=1}^{n} \xi_i$$

$$\text{s.t. } 1 - y_i x_i^T w \leq \xi_i$$

$$\xi_i \geq 0$$

Maximum Likelihood

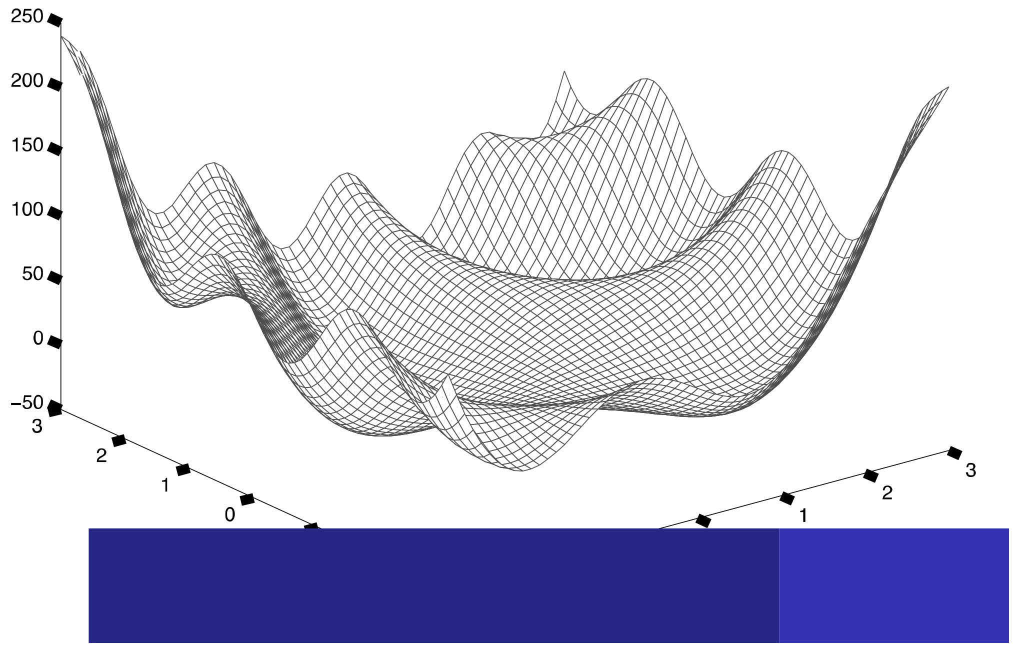
$$\arg\max_{\theta} \sum_{i=1}^{n} \log\, p_\theta(x_i)$$

K-Means

$$\arg \min_{\mu_1,\mu_2,\ldots,\mu_k} J(\mu) = \sum_{j=1}^{k} \sum_{i \in C_j} ||x_i - \mu_j||^2$$

Prefer Convex Problems

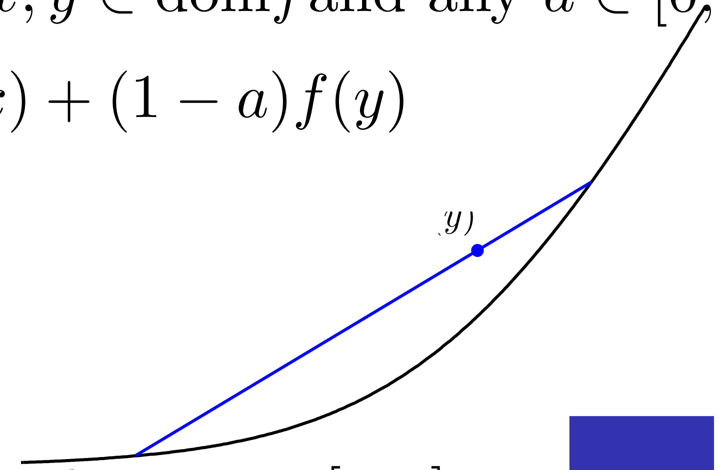
Local (non global) minima and maxima:



Convex Functions and Sets

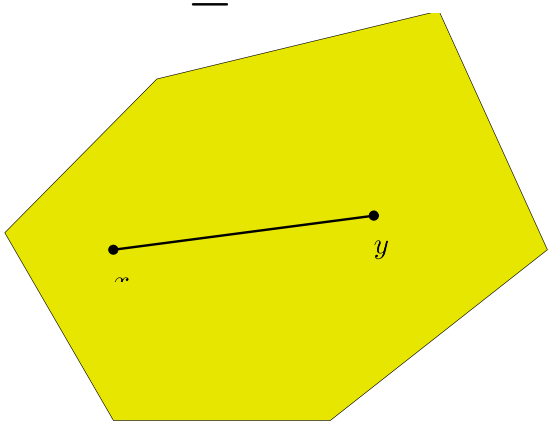
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \mathrm{dom} f$ and any $a \in [0, 1]$,

$$f(ax + (1-a)y) \leq af(x) + (1-a)f(y)$$



A set $C \subseteq \mathbb{R}^n$ is convex if for $x, y \in C$ and any $a \in [0, 1]$,

$$ax + (1-a)y \in C$$



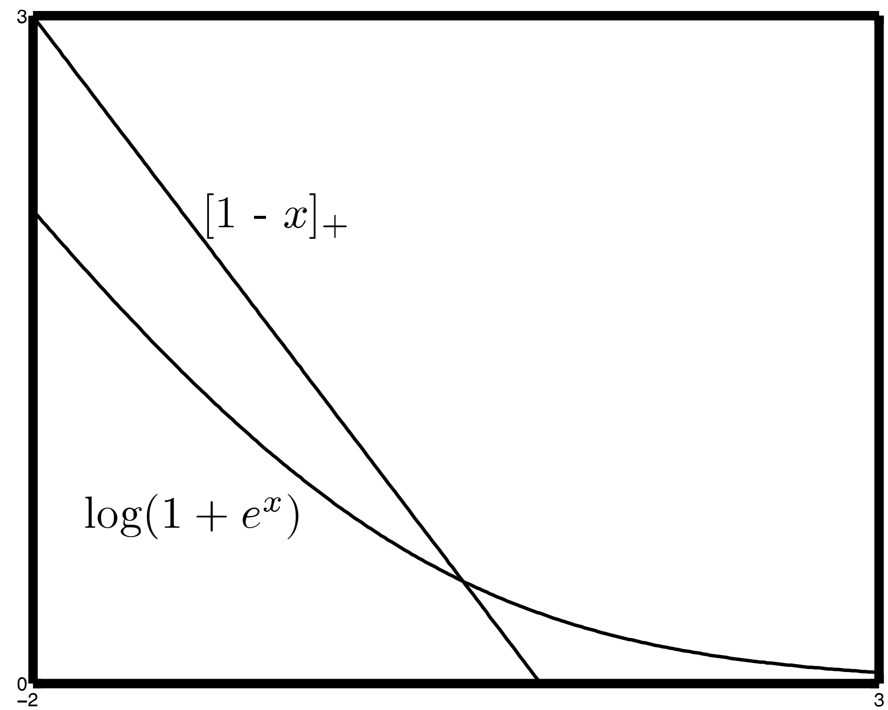
Important Convex Functions

SVM loss:

$$f(w) = \left[1 - y_i x_i^T w\right]_+$$

Binary logistic loss:

$$f(w) = \log\left(1 + \exp(-y_i x_i^T w)\right)$$



Convex Optimization Problem

$$\begin{aligned} \underset{x}{\text{minimize}} \ & f_0(x) && \text{(Convex function)} \\ \text{s.t.} \quad & f_i(x) \leq 0 && \text{(Convex sets)} \\ & h_j(x) = 0 && \text{(Affine)} \end{aligned}$$

Lagrangian Dual

Start with optimization problem:

$$\begin{aligned} \underset{x}{\text{minimize}}\ & f_0(x) \\ \text{s.t.}\ \ & f_i(x) \leq 0,\ i = \{1, \ldots, k\} \\ & h_j(x) = 0,\ j = \{1, \ldots, l\} \end{aligned}$$

Form *Lagrangian* using Lagrange multipliers $\lambda_i \geq 0$, $\nu_i \in \mathbb{R}$

$$\mathcal{L}(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{l} \nu_j h_j(x)$$

Form *dual function*

$$g(\lambda, \nu) = \inf_x \mathcal{L}(x, \lambda, \nu) = \inf_x \left\{ f_0(x) + \sum_{i=1}^{k} \lambda_i f_i(x) + \sum_{j=1}^{l} \nu_j h_j(x) \right\}$$

First Order Methods:

Gradient Descent

Newton’s Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Gradient Descent

The simplest algorithm in the world (almost). Goal:

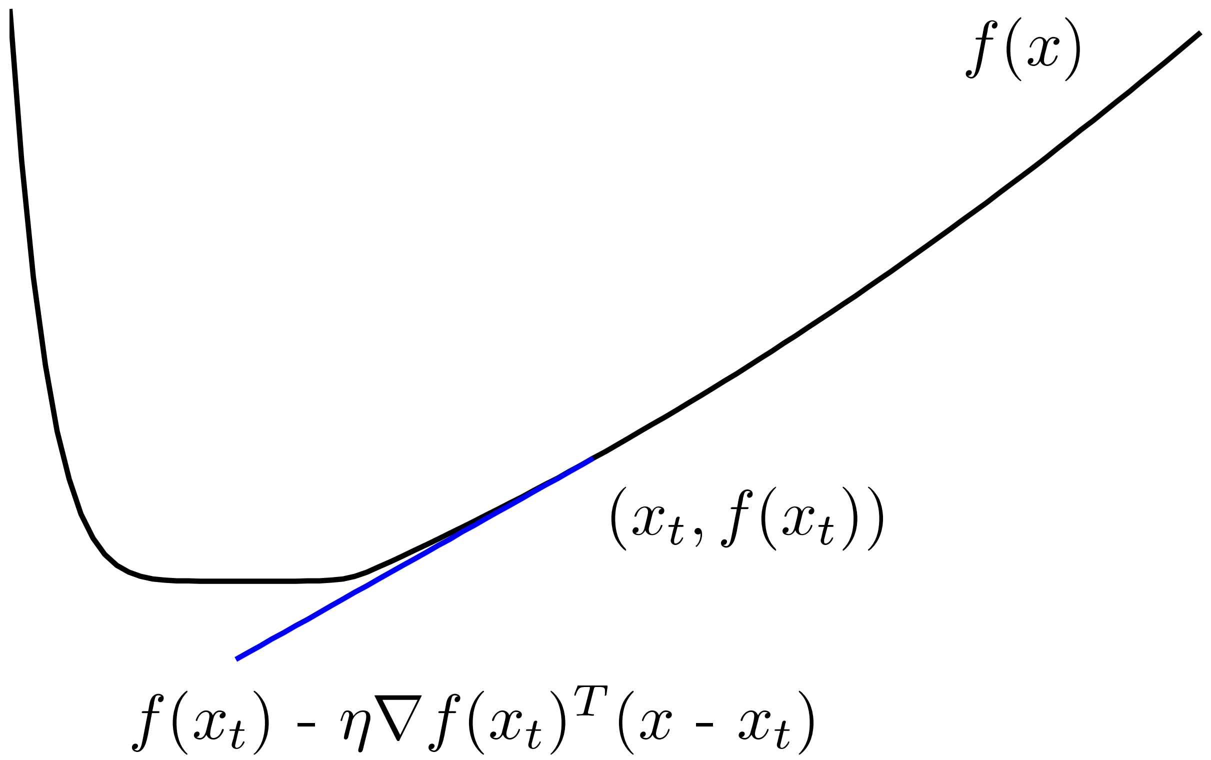
$$\underset{x}{\text{minimize}} \;\; f(x)$$

Just iterate

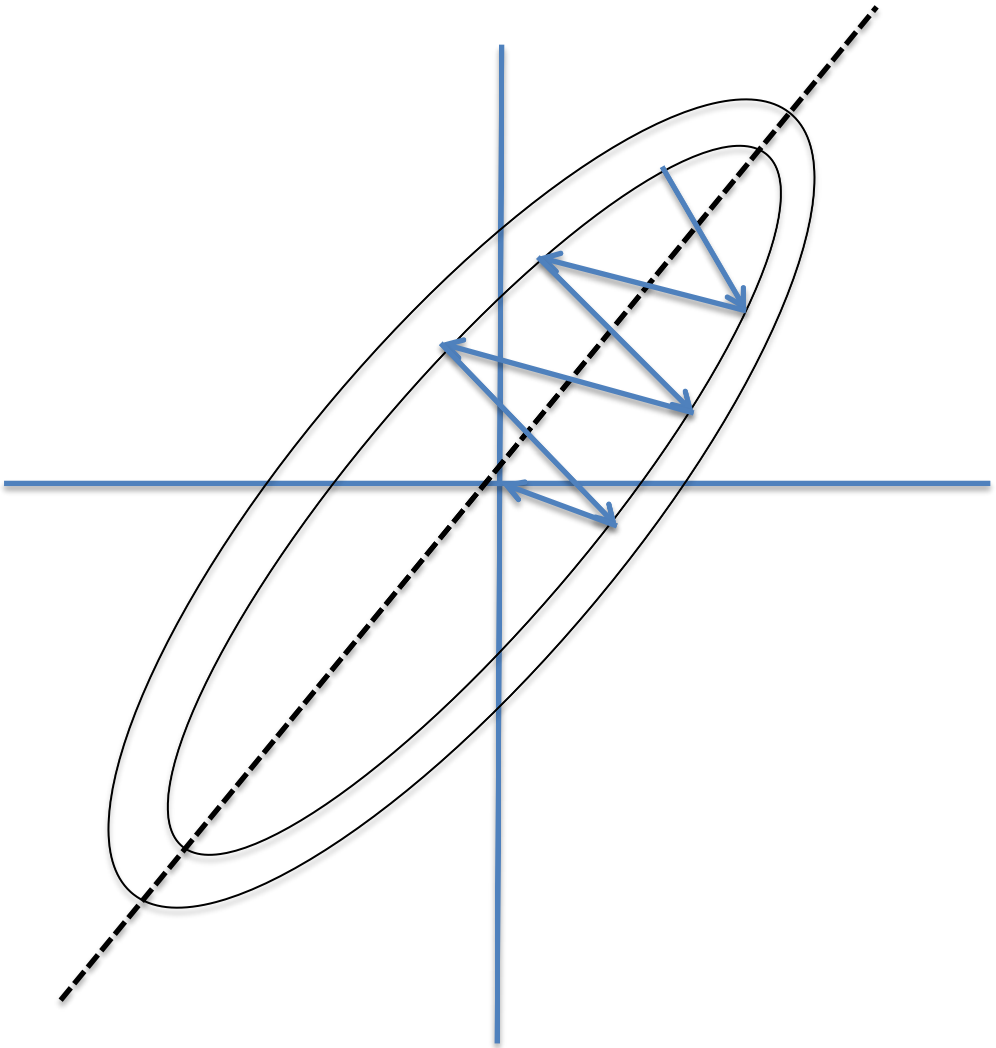
$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

where η_t is stepsize.

Single Step Illustration



Full Gradient Descent Illustration



First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton’s Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Newton's Method

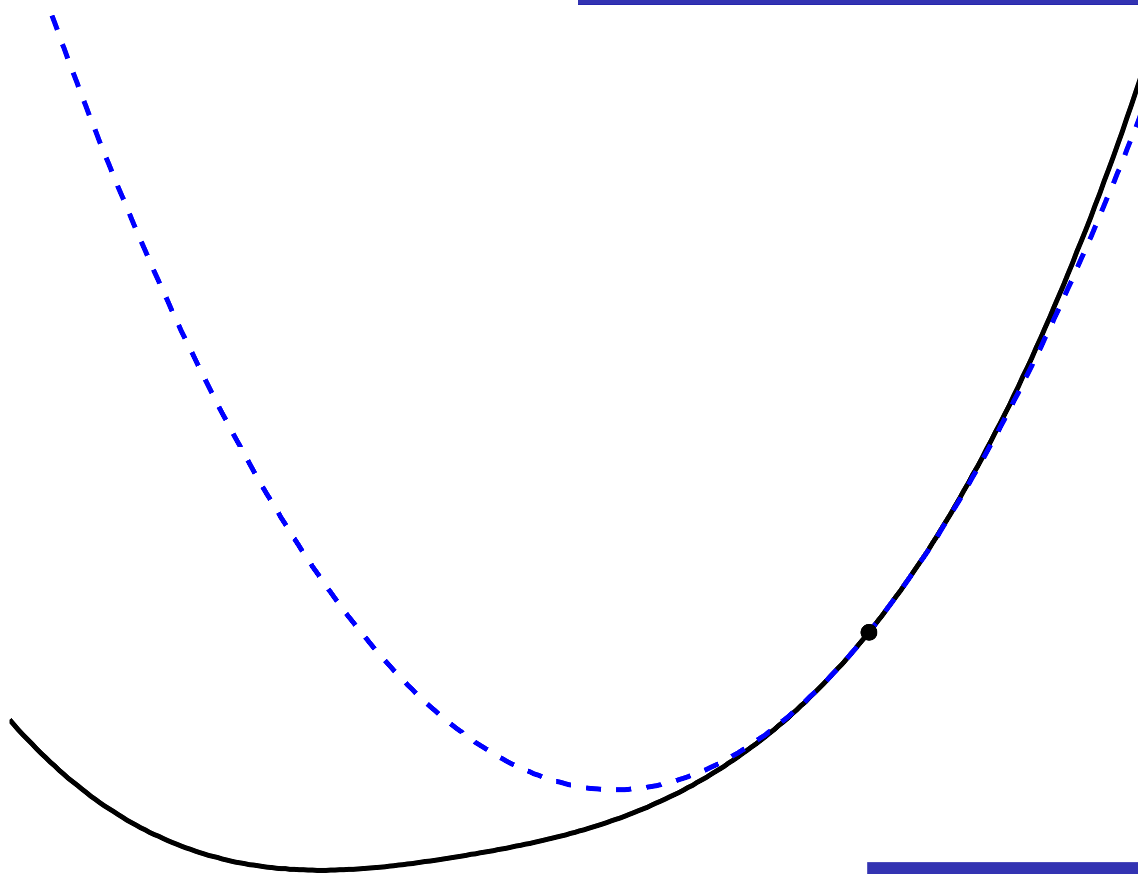
Idea: use a second-order approximation to function.

$$f(x + \Delta x) \approx f(x) + \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \Delta x$$

Choose Δx to minimize above:

$$\Delta x = - \underbrace{\left[\nabla^2 f(x)\right]^{-1}}_{\text{Inverse Hessian}} \underbrace{\nabla f(x)}_{\text{Gradient}}$$

Newton’s Method Picture



First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton’s Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

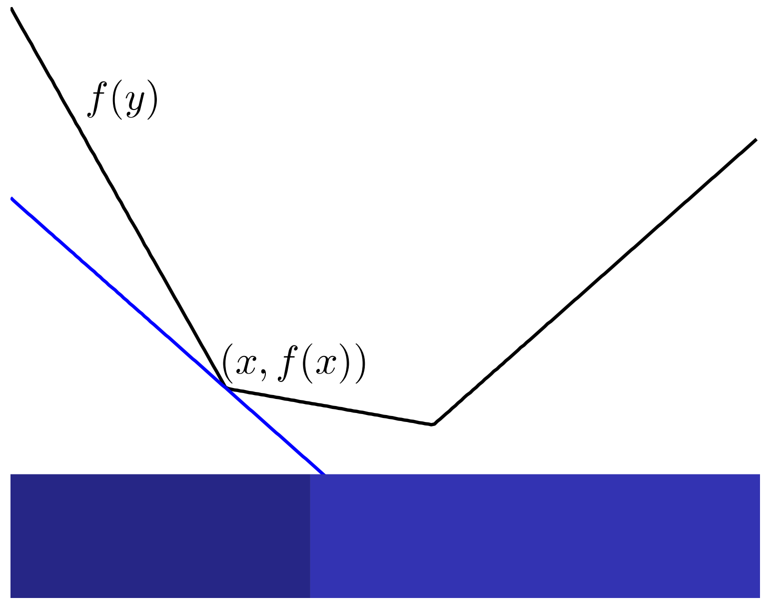
Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Subgradient Descent Motivation

Lots of non-differentiable convex functions used in machine learning:



The *subgradient set*, or subdifferential set, $\partial f(x)$ of f at x is

$$\partial f(x) = \left\{ g : f(y) \geq f(x) + g^T (y - x) \;\; \text{for all } y \right\} .$$

Subgradient Descent – Algorithm

Really, the simplest algorithm in the world. Goal:

$$\underset{x}{\text{minimize}} \;\; f(x)$$

Just iterate

$$x_{t+1} = x_t - \eta_t g_t$$

where η_t is a stepsize, $g_t \in \partial f(x_t)$.

First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton’s Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Online learning and optimization

- Goal of machine learning :
  - Minimize expected loss

$$\min_h L(h) = \mathbf{E}\left[\mathrm{loss}(h(x), y)\right]$$

  given samples $(x_i, y_i)\ i = 1, 2...m$

- This is Stochastic Optimization
  - Assume loss function is convex

Batch (sub)gradient descent for ML

- Process all examples together in each step

$$w^{(k+1)} \leftarrow w^{(k)} - \eta_t \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial L(w, x_i, y_i)}{\partial w} \right)$$

where L is the regularized loss function

- Entire training set examined at each step
- Very slow when *n* is very large

Stochastic (sub)gradient descent

- "Optimize" one example at a time
- Choose examples randomly (or reorder and choose in order)
  - Learning representative of example distribution

for $i = 1$ to n:

$$w^{(k+1)} \leftarrow w^{(k)} - \eta_t \frac{\partial L(w, x_i, y_i)}{\partial w}$$

where L is the regularized loss function

Stochastic (sub)gradient descent

for $i = 1$ to n:

$$w^{(k+1)} \leftarrow w^{(k)} - \eta_t \frac{\partial L(w, x_i, y_i)}{\partial w}$$

where L is the regularized loss function

- Equivalent to online learning (the weight vector *w* changes with every example)
- Convergence guaranteed for convex functions (to local minimum)

Hybrid!

- Stochastic – 1 example per iteration
- Batch – All the examples!
- Sample Average Approximation (SAA):
  - Sample *m* examples at each step and perform SGD on them
- Allows for parallelization, but choice of *m* based on heuristics

SGD - Issues

- Convergence very sensitive to learning rate (η_t) (oscillations near solution due to probabilistic nature of sampling)
  – Might need to decrease with time to ensure the algorithm converges eventually
- Basically – SGD good for machine learning with large data sets!

First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Problem Formulation

At x, want Δx. At x have model of function q_k

$$q_k = \nabla f(x)^T \Delta x + \frac{1}{2} \Delta x^T \nabla^2 f(x) \nabla x$$

I trust the model in the region of size d_k

Need to:

-Generate new point

-Generate trust region around new point

New Points

Update x and Δx by checking

$$\rho = \frac{f(x) - f(x + \Delta x)}{q(x) - q(x + \Delta x)}$$

If $\rho \sim 1$ good model, make d_{k+1} bigger

If ρ small good model, make $d_{k+1} = d_k$

If $\rho \leq 1$ reject step, make d_k smaller

Limited Memory Quasi-Newton Methods

use techniques to obtain approximate inverse Hessian H_k
and update it to H_{k+1}

One of the most popular such methods is LBFGS

LBFGS restricts the update to use only m vectors from previous iterations.

Limited Memory BFGS

Given $\mathbf{w}^0, H^0$ and a small integer m.

For $k = 0, 1, \ldots$

- If $\nabla f(\mathbf{w}^k) = \mathbf{0}$, stop.
- Using m vectors from previous iterations to calculate $H_k \nabla f(\mathbf{w}^k)$, where H_k is an approximate inverse Hessian.
- Search α_k so that
  $$f(\mathbf{w}^k - \alpha H_k \nabla f(\mathbf{w}^k))$$
  satisfies certain sufficient decrease conditions.
- Update H_k to H_{k+1}.

First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton’s Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Coordinate descent

- Minimize along each coordinate direction in turn. Repeat till minimum is found
  - One complete cycle of coordinate descent is the same as gradient descent
- In some cases, analytical expressions available:
  - Example: Dual form of SVM!
- Otherwise, numerical methods needed for each iteration

Dual coordinate descent

- Coordinate descent applied to the dual problem
- Commonly used to solve the dual problem for SVMs
  - Allows for application of the Kernel trick
  - Coordinate descent for optimization
- In this paper: Dual logistic regression and optimization using coordinate descent

Dual form of SVM

- SVM

$$\min_w P(w) = C\sum_{i=1}^{l} \max(1 - y_i w^T x_i, 0) + \frac{1}{2} w^T w$$

- Dual form

$$\min_\alpha D(\alpha) = \tfrac{1}{2}\alpha^T Q\alpha - \sum_{i=1}^{l} \alpha_i$$

$$\text{Subject to } 0 \leq \alpha \leq C$$

Dual form of LR

- LR:

$$\textbf{Minimize:} \quad P(w) = C\sum_{i=1}^{n}\log\left(1+e^{-y_i w^T x_i}\right) + \frac{1}{2}w^T w$$

- Dual form (we let $w = \sum_{i=1}^{n} \alpha_i y_i x_i$)

$$\min_\alpha D(\alpha) = \tfrac{1}{2}\alpha^T Q\alpha + \sum_{i:\alpha_i>0} \alpha_i \log \alpha_i + \sum_{i:\alpha_i<C} (C-\alpha_i)\log(C-\alpha_i)$$

$$\text{Subject to } 0 \le \alpha \le C, \text{ and } Q_{ij} = y_i y_j x_i^T x_j$$

Coordinate descent for dual LR

$$\min_\alpha D(\alpha) = \frac{1}{2}\alpha^T Q\alpha + \sum_{i:\alpha_i>0} \alpha_i \log \alpha_i + \sum_{i:\alpha_i<C} (C-\alpha_i)\log(C-\alpha_i)$$

Subject to $0 \le \alpha \le C$, and $Q_{ij} = y_i y_j x_i^T x_j$

- Along each coordinate direction:

$$\min_z g(z) = \frac{a}{2}z^2 + bz + (c_1+z)\log(c_1+z) + (c_2-z)\log(c_2-z)$$

Subject to: $-c_1 \le z \le c_2$

where $c_1 = \alpha_i, c_2 = C - \alpha_i, a = Q_{ii}$ and $b = (Q\alpha)_i$

Coordinate descent for dual LR

$\min_z g(z) = \frac{a}{2}z^2 + bz + (c_1 + z)\log(c_1 + z) + (c_2 - z)\log(c_2 - z)$
Subject to: $-c_1 \leq z \leq c_2$
where $c_1 = \alpha_i, c_2 = C - \alpha_i, a = Q_{ii}$ and $b = (Q\alpha)_i$

- No analytical expression available
  - Use numerical optimization (Newton's method/bisection method/BFGS/...)to iterate along each direction
- Beware of log!

Coordinate descent for dual ME

- Maximum Entropy (ME) is extension of LR to multi-class problems
  - In each iteartion, solve in two levels:
    - Outer level – Consider block of variables at a time
      - Each block has all labels and **one** example
    - Inner level – Subproblem solved by dual coordinate descent
- Can also be solved similar to online CRF (exponentiated gradient methods)

First Order Methods:

Gradient Descent

Newton's Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

First Order Methods:

Gradient Descent

Newton’s Method

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Subgradient Descent

Introduction to Convex Optimization for Machine Learning, John Duchi, UC Berkeley, Tutorial, 2009

Stochastic Gradient Descent

Stochastic Optimization for Machine Learning, Nathan Srebro and Ambuj Tewari, presented at ICML'10

Trust Regions

Trust Region Newton method for large-scale logistic regression, C.-J. Lin, R. C. Weng, and S. S. Keerthi, Journal of Machine Learning Research, 2008

Dual Coordinate Descent

Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

Linear Classification

Recent Advances of Large-scale linear classification, G.-X. Yuan, C.-H. Ho, and C.-J. Lin. Proceedings of the IEEE, 100(2012)

Large scale linear classification

- NLP (usually) has large number of features, examples

- Nonlinear classifiers (including kernel methods) more accurate, but slow

Large scale linear classification

- Linear classifiers less accurate, but at least an order of magnitude faster
  - Loss in accuracy lower with increase in number of examples
- Speed usually dependent on more than algorithm order
  - Memory/disk capacity
  - Parallelizability

Large scale linear classification

- Choice of optimization method depends on:
  - Data property
    - Number of examples, features
    - Sparsity
  - Formulation of problem
    - Differentiability
    - Convergence properties
  - Primal vs dual
  - Low order vs high order methods

Comparison of performance

- Performance gap goes down with increase in number of features
- Training, testing time for linear classifiers is much faster

Data set	#instances		#features	Linear			Nonlinear (kernel)			Accuracy difference to nonlinear
				Time (s)		Testing accuracy	Time (s)		Testing accuracy	
	Training	Testing		Training	Testing		Training	Testing		
cod-RNA	59,535	271,617	8	3.1	0.05	70.71	80.2	126.02	96.67	-25.96
ijcnn1	49,990	91,701	22	1.7	0.01	92.21	26.8	20.29	98.69	-6.48
covtype	464,810	116,202	54	1.5	0.03	76.37	46,695.8	1,131.20	96.11	-19.74
webspam	280,000	70,000	254	26.8	0.04	93.35	15,681.8	853.34	99.26	-5.91
MNIST38	11,982	1,984	752	0.2	0.01	96.82	38.1	5.61	99.70	-2.88
real-sim	57,848	14,461	20,958	0.3	0.01	97.44	938.3	81.94	97.82	-0.38
rcv1	20,242	677,399	47,236	0.1	0.43	96.26	108.0	3,259.46	96.50	-0.24
astro-physic	49,896	12,473	99,757	0.3	0.01	97.09	735.7	111.59	97.31	-0.22
yahoo-japan	140,963	35,240	832,026	3.3	0.03	92.63	20,955.2	1,890.83	93.31	-0.68
news20	15,997	3,999	1,355,191	1.2	0.03	96.95	383.2	100.38	96.90	0.05

References:

[1] S.V Rozhkova, O.N. Imas, *Methods of optimization*", Tomsk Polytechnic University 2004.

[2] Thomas S. Ferguson, *Linear Programming – A concise Introduction.*

[3] James K. Strayer, Linear Programming and Applications, (1989) Springer-Verlag.

[4] *Linear Programming: Geometric Approach,* Willey Publications.

[5] *Math 407A: Linear Optimization,* Lecture 4: LP Standard Form Math Dept., University of Washington.

[6] Mark A. Schulze, *Linear Programming for Optimization*, Perceptive Scientific Instruments, Inc.

[7] Optimization Methods in Management sciences / Operation Research, Spring 2013, MIT OpenCourseWare.

[8] R. T. Rockafellar, FUNDAMENTALS OF OPTIMIZATION 2007 Lecture Notes, Dept. of Mathematics University of Washington.

12. Kočegurova E.A. Theory and optimization techniques [electronic resource]: the manual.-Tomsk: Publishing House of TPU, 2013. The access Scheme: http://www.lib.tpu.ru/fulltext2/m/2013/m234.pdf
13. Fedunec, Nina Ivanovna. Optimization techniques: A training manual for high schools/N. I. Fedunec, Y. V. Chernikov; Moscow State Mining University (MSMU).-Moscow: IZD-vo MSMU, 2009.-375 p.
14. Victor A. Goncharov. Optimization techniques [electronic resource]: tutorial/V.A. Goncharov; National research University, Moscow Institute of electronic technology (ELECTRONIC TECHNOLOGY).- The access Scheme: http://www. lib . tpu . ru / fulltext 2/ (m) //2014 FN / fn -01. pdf
15. D.l.Horvata, V.A.Gluharev. Optimization techniques: Tutorial and workshop. Moscow State University (MGU).-3 ed., Corr. and additional charge. - Moscow: Harvard Business Press, 2014. — p 367.
16. Sofieva, Julia. Conditional optimization: methods and objectives/Y. M. Sofieva, A. M. Zirlin. -Ed. 2. - Moscow: Librokom, 2012. -143 p.