CS639:
Data Management for Data Science

Lecture 20: Optimization/Gradient Descent

Theodoros Rekatsinas

Today

1. Optimization

2. Gradient Descent

What is Optimization?

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

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

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

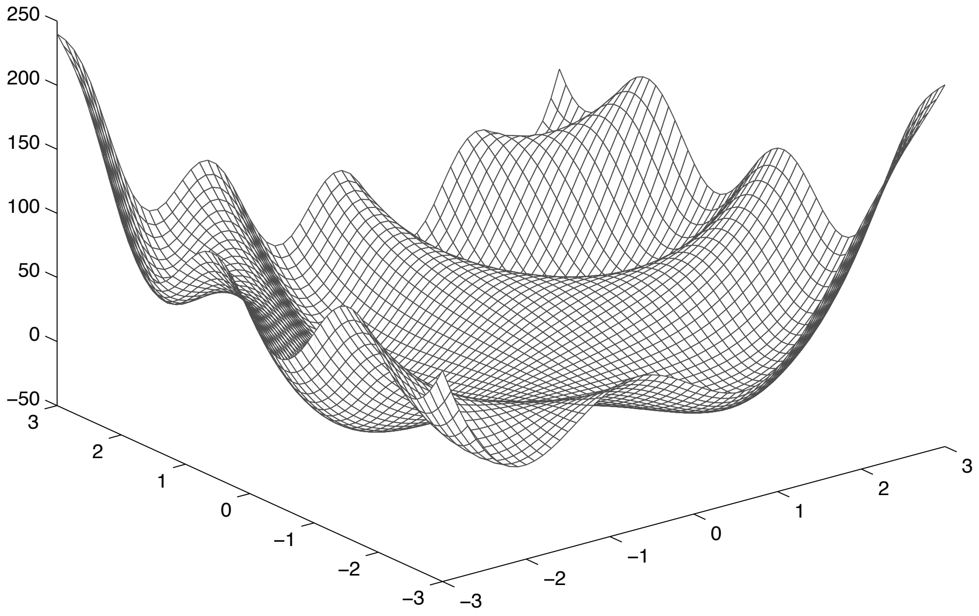
$$\arg\max_{\theta} \sum_{i=1}^{n} \log\, p_\theta(x_i)$$

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

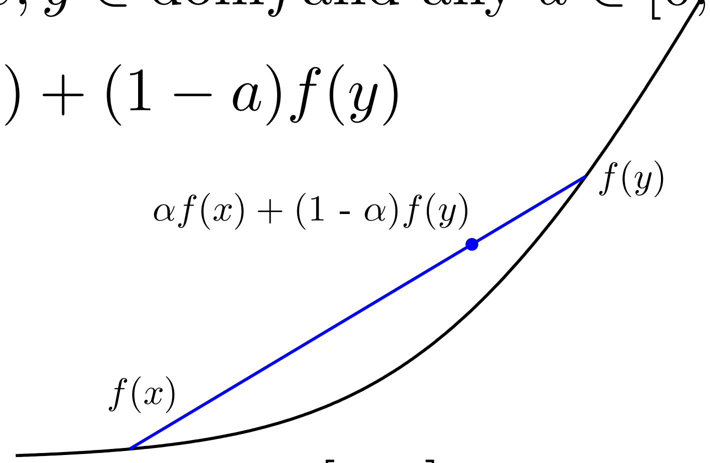
Local (non global) minima and maxima:



Convex Functions and Sets

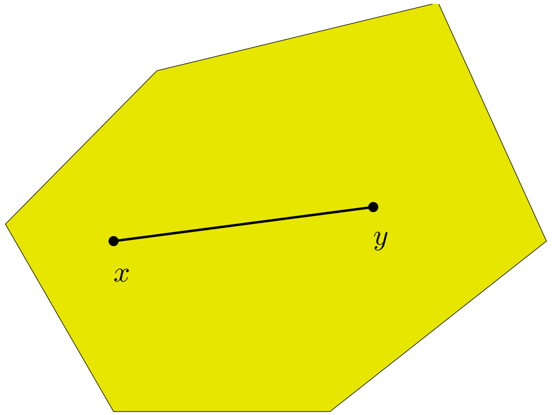
A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for $x, y \in \mathrm{dom} f$and any $a \in [0, 1]$,

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



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

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



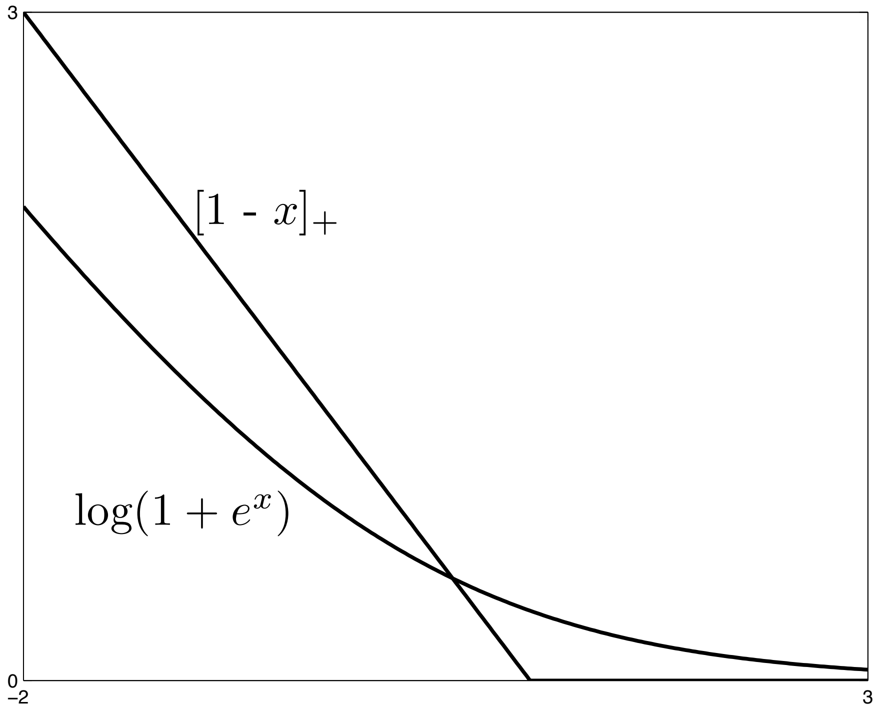
Important Convex Functions

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:

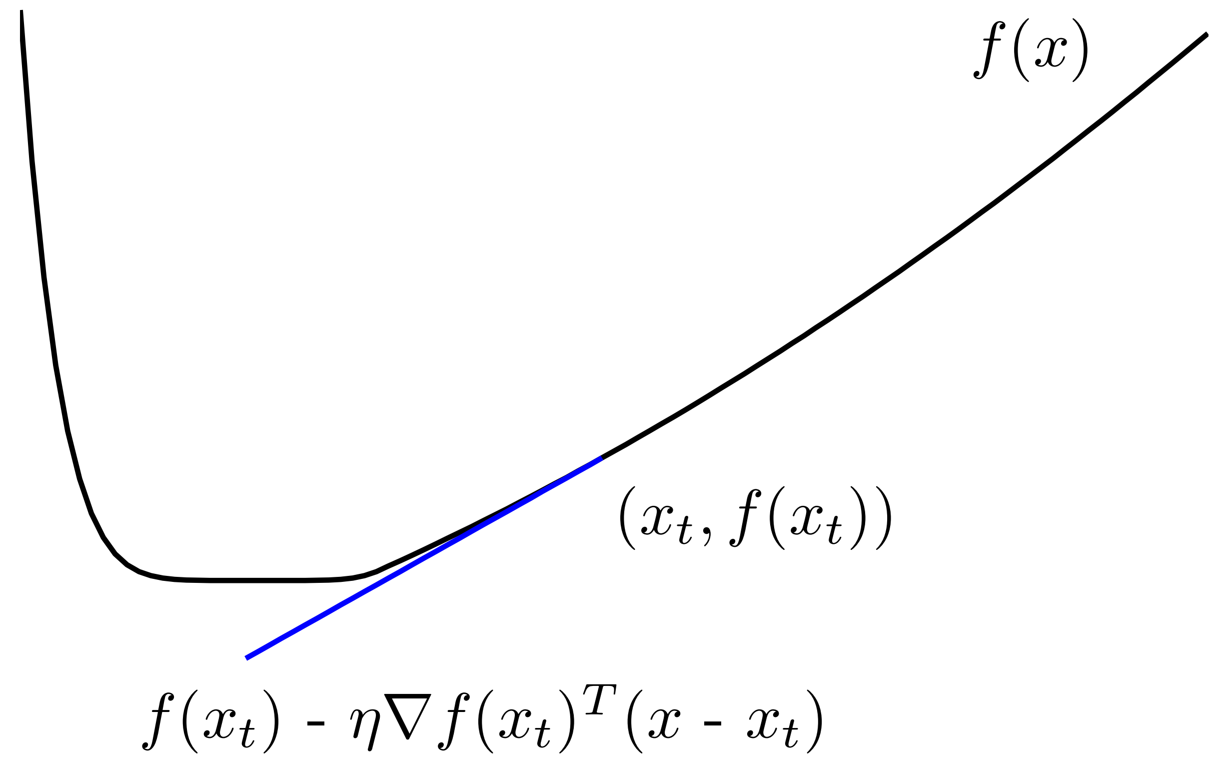
$$\underset{x}{\text{minimize}} \;\; f(x)$$

Just iterate

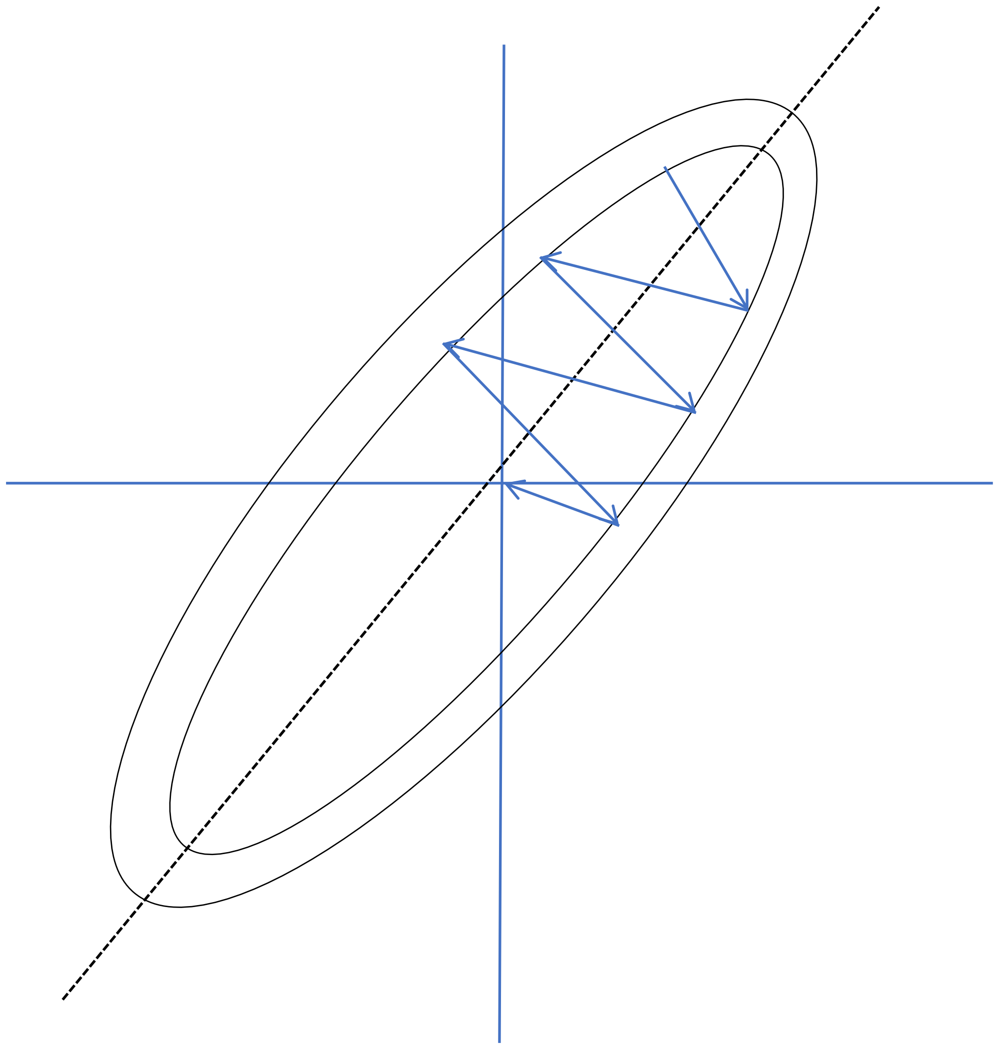
$$x_{t+1} = x_t - \eta_t \nabla f(x_t)$$

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

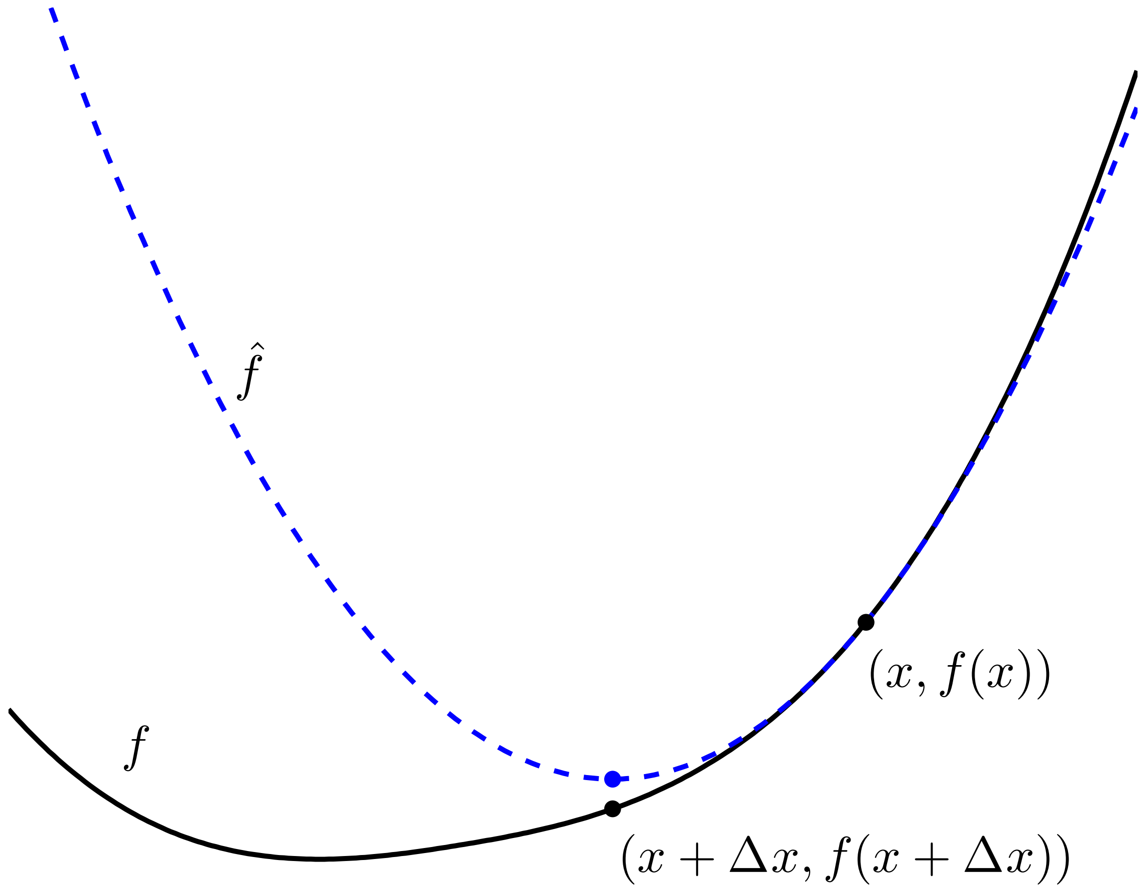
Idea: use a second-order approximation to function.

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

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



$\hat{f}$ is 2^{nd}-order approximation, f is true function.

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

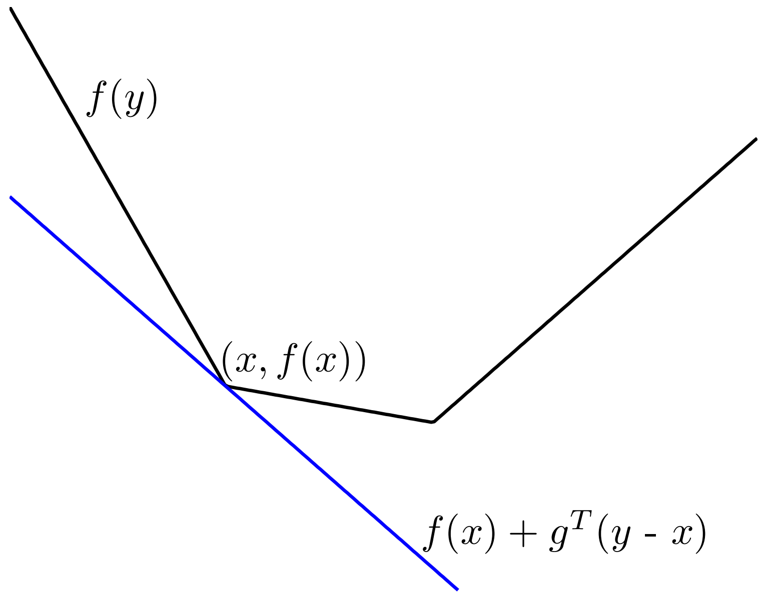
Dual Coordinate Descent Methods for logistic regression and maximum entropy models, H.-F. Yu, F.-L. Huang, and C.-J. Lin, . Machine Learning Journal, 2011

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

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

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

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

$$\text{where } L \text{ is the regularized loss function}$$

- 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
  - ( ) (oscillations near solution due to probabilistic nature of sampling) η_t
  - 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)

Statistical Learning Crash Course

Staggering amount of machine learning/stats can be written as:

$$\min_x \sum_{i=1}^{N} f(x, y_i)$$

N (number of y_i's, data) typically in the billions
Ex: Classification, Recommendation, Deep Learning.

De facto iteration to solve large-scale problems: **SGD**.

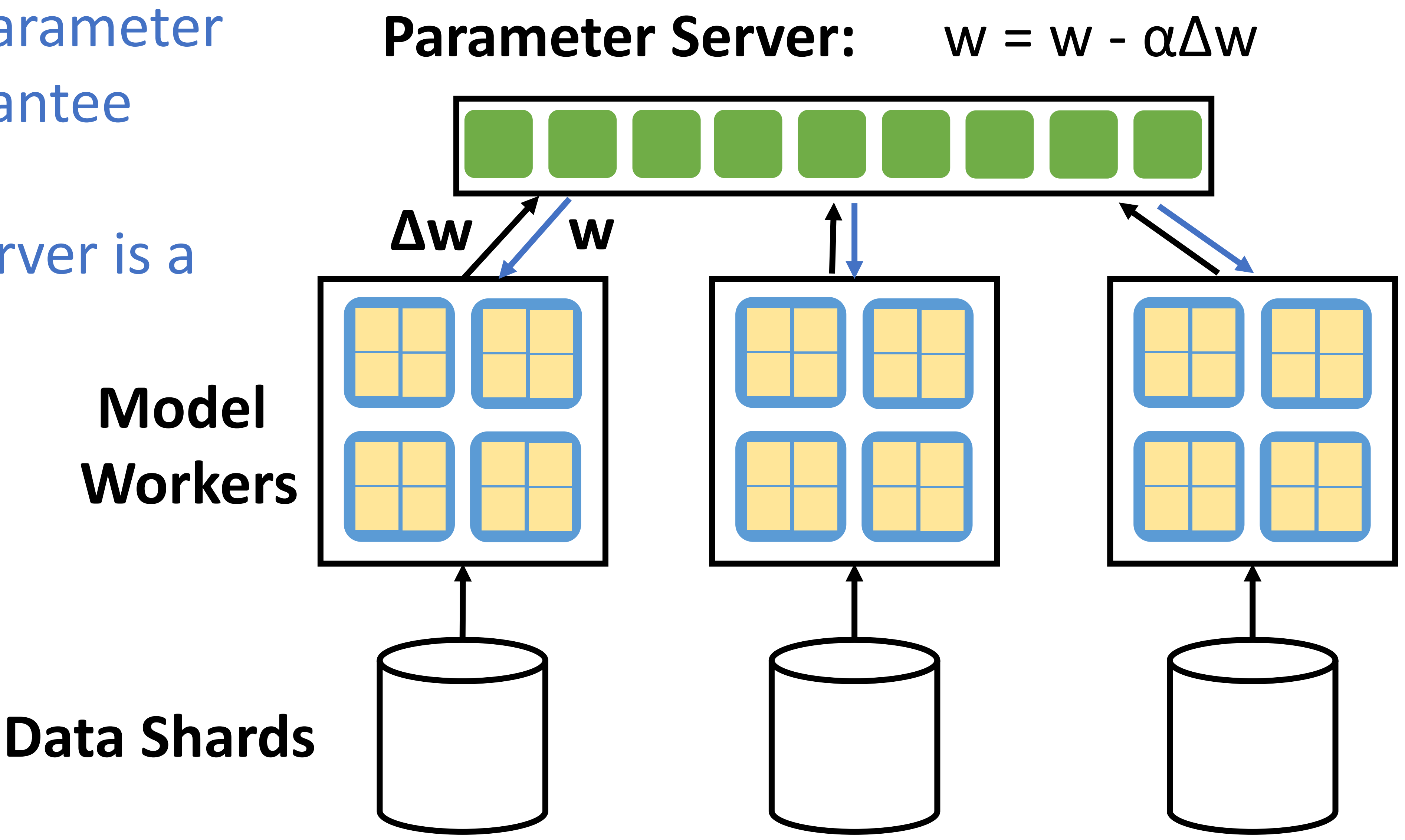
$$\boxed{x^{k+1} = x^k - \alpha N \nabla f(x^k, y_j)}$$

Billions of tiny iterations

Select one term, j, and estimate gradient.

Parallel SGD (Centralized)

- Centralized parameter updates guarantee convergence
- Parameter Server is a bottleneck



Parallel SGD (HogWild! - asynchronous)

Data Systems Perspective of SGD

$$x^{k+1} = x^k - \alpha N \nabla f(x^k, y_j)$$

Insane conflicts: Billions of tiny jobs (~100 instructions), RW conflicts on x

Multiple workers need to **communicate**!

HogWild!: For sparse convex models (e.g., logistic regression) run without locks; SGD still converges (answer is statistically correct)