Optimization in Machine Learning

Tong Zhang

Rutgers University

Topics

- Gradient Descent
- Proximal Projection Method
- Coordinate Descent
- Convex Duality and Dual Coordinate Descent
- LBFGS

Supervised Learning

- Training data: (X_i, Y_i) $(i = 1, \ldots, n)$
- Example: linear prediction function $w^T x$
- Training algorithm: SVM

$$\hat{w} = \arg\min_w \left[n^{-1} \sum_{i=1}^{n} (1 - w^T X_i Y_i)_+ + \lambda w^T w \right].$$

- This is an optimization problem: how to find w?

Unconstrained Optimization

Consider a general unconstrained optimization problem:

$$w_* = \arg\min_w f(w),$$

How to find the optimal solution?

- global solution: w such that $f(w) \leq f(w')$ for all w'.
- local solution: w such that $f(w) \leq f(w')$ when w' is close to w.
- global solution is local solution but not necessarily vice versa.
- local optimal (and thus global optimal) solution satisfies $\nabla f(w) = 0$.
- for convex problems: local and global solutions are the same.

Gradient Descent

$$w_k = w_{k-1} - \eta_k \nabla f(w_{k-1}).$$

How fast does this method converge to the optimal solution?

- General result: converge to local minimum under suitable conditions.
- What's the convergence rate?

Gradient Descent

$$w_k = w_{k-1} - \eta_k \nabla f(w_{k-1}).$$

How fast does this method converge to the optimal solution?

- General result: converge to local minimum under suitable conditions.
- What's the convergence rate?
- Answer: depends on conditions of $f(\cdot)$.
- This lecture focuses on convex problems.

Convexity

For all $\alpha \in [0, 1]$, we have

$$f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha) f(x').$$

Convexity

For all $\alpha \in [0, 1]$, we have

$$f(\alpha x + (1 - \alpha)x') \leq \alpha f(x) + (1 - \alpha) f(x').$$

A subgradient $\nabla f(x_0)$ at x_0 satisfies:

$$f(x) \geq f(x_0) + (x - x_0)^\top \nabla f(x_0)$$

- Generalize gradient for functions
- Subgradient v_0 is not necessarily unique:
  - $f(x) = |x|$ ($x \in R$)
  - at $x_0 = 0$: any $v_0 \in [-1, 1]$ satisfies the requirement (thus a subgradient)

In the following we assume subgradient always exists

Common Conditions of Objective Function

Convexity:

$$f(x') - f(x) - \nabla f(x)^T (x' - x) \geq 0$$

- Nonsmooth: first order derivative may be discontinuous: e.g. hinge loss or L_1 regularization
- Smooth: first order derivative is Lipschitz or second order derivative is bounded:

$$f(x') - f(x) - \nabla f(x)^T (x' - x) \leq \frac{L}{2} \|x' - x\|_2^2$$

- Strongly Convex:

$$f(x') - f(x) - \nabla f(x)^T (x' - x) \geq \frac{\mu}{2} \|x' - x\|_2^2$$

  Convex is when the above satisfied with $\mu = 0$.

A function can be strongly convex and nonsmooth: $f(x) = x^2 + |x|$.

Results

- Smooth and strongly convex: gradient descent with a sufficiently small constant η_k has linear (or geometric) convergence:

$$f(w_k) - f(w_*) = O(\gamma^k)$$

  for some $\gamma < 1$.
- Smooth but not strongly convex:

$$f(w_k) - f(w_*) = O(1/k),$$

  with learning rate $\eta_k = O(1/k)$.
- Nonsmooth:

$$f(\tilde{w}_k) - f(w_*) = O(1/\sqrt{k}),$$

  for $\eta_k = O(1/\sqrt{k})$ and $\tilde{w}_k = k^{-1} \sum_{j=1}^{n} w_k$.

The learning rate can be tuned with line search.

Reformulation of Gradient Descent

Gradient descent can be derived from:

$$w_k = \arg\min_w Q_k(w)$$

$$Q_k(w) := f(w_{k-1}) + \nabla f(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2\eta_k} \|w - w_{k-1}\|_2^2$$

Reformulation of Gradient Descent

Gradient descent can be derived from:

$$w_k = \arg\min_w Q_k(w)$$

$$Q_k(w) := f(w_{k-1}) + \nabla f(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2\eta_k} \|w - w_{k-1}\|_2^2$$

Key properties: assume smoothness for simplicity and $1/\eta_k \geq L$ (smoothness parameter of f).

- $Q_k(w_{k-1}) = f(w_{k-1})$
- $Q_k(w) \geq f(w)$
- $Q_k(w)$ is easy to optimize

Reformulation of Gradient Descent

Gradient descent can be derived from:

$$w_k = \arg\min_w Q_k(w)$$

$$Q_k(w) := f(w_{k-1}) + \nabla f(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2\eta_k} \|w - w_{k-1}\|_2^2$$

Key properties: assume smoothness for simplicity and $1/\eta_k \geq L$ (smoothness parameter of f).

- $Q_k(w_{k-1}) = f(w_{k-1})$
- $Q_k(w) \geq f(w)$
- $Q_k(w)$ is easy to optimize

Consequence: minimize $Q_k(w)$ reduces objective value of $f(w)$: $f(w_{k-1}) - f(w_k) \geq Q_k(w_{k-1}) - Q_k(w_k)$.

This idea can be be generalized to other convex upper bound of $f(w)$.

Proximal Gradient Method

Assume

$$f(w) = \phi(w) + g(w),$$

then we may consider the following upper bound of $f(w)$

$$Q_k(w) := \phi(w_{k-1}) + \nabla\phi(w_{k-1})^T(w - w_{k-1}) + \frac{1}{2\eta_k}\|w - w_{k-1}\|_2^2 + g(w),$$

with $1/\eta_k$ larger than the smoothness parameter of ϕ. Then solve for

$$w_k = \arg\min_w Q_k(w).$$

We assume that this minimization problem is easy.

- generalization of gradient descent called proximal gradient descent.
- useful when $g(w)$ is a simple nonsmooth function such as L_1 regularization $g(w) = \lambda\|w\|_1$.

Example: L_1 regularization

$$f(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1.$$

For example, $\phi(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$ and $g(w) = \lambda \|w\|_1$. Then

$$Q_k := \phi(w_{k-1}) + \nabla \phi(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2\eta_k} \|w - w_{k-1}\|_2^2 + \lambda \|w\|_1.$$

Example: L_1 regularization

$$f(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_1.$$

For example, $\phi(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2$ and $g(w) = \lambda \|w\|_1$. Then

$$Q_k := \phi(w_{k-1}) + \nabla \phi(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2\eta_k} \|w - w_{k-1}\|_2^2 + \lambda \|w\|_1.$$

Solution is

$$w_k = \mathrm{trunc}(w_{k-1} - \eta_k \nabla \phi(w_{k-1})),$$

where

$$\begin{aligned} \mathrm{trunc}([u_1, \ldots, u_d]) &= [\mathrm{trunc}(u_j)]_{j=1,\ldots,d} \\ \mathrm{trunc}(u_j) &= \mathrm{sign}(u_j)(|u_j| - \lambda \eta_k)_+ \end{aligned}$$

Property of Proximal Gradient

Smoothness depending on ϕ rather than f (can tolerate nonsmooth g)
Convergence similar to gradient descent:

- if ϕ is smooth and f is strongly convex: convergence is linear
- if ϕ is smooth but not strongly convex: convergence is $1/k$.
- if ϕ is not smooth: convergence is $1/\sqrt{k}$.

Nesterov's Accelerated Gradient (one version)

Procedure:

- Pick $\eta_1, \eta_2, \ldots \geq 0$
- Pick $w_1 = y_1 = z_0$, then
- Define $\alpha_0 = 0$ and $\alpha_i^{-2} - \alpha_i^{-1} = \alpha_{i-1}^{-2}$ for $i \geq 1$
  (may also set $\alpha_i = (1 + i/2)^{-1}$)
- Iterate for $i = 1, 2, \ldots, T$:

$$z_i = \arg\min_z \left[g(z) + \frac{1}{2\eta_i}\|z\|_2^2 - (\eta_i^{-1} z_{i-1} - \alpha_i^{-1}\nabla\phi(y_i))^\top z \right],$$

$$w_i = (1 - \alpha_{i-1}) w_{i-1} + \alpha_{i-1} z_i$$

$$y_{i+1} = (1 - \alpha_i) w_i + \alpha_i z_i$$

Advantage: faster convergence of $1/k^2$ for smooth ϕ
Disadvantage:

- for smooth and strongly convex f: algorithm has to be modified to achieve geometric convergence
- modification depends on strong convexity parameter μ.

Beyond First Order Method: LBFGS (high level view)

- Recall gradient descent: successive minimization of

$$Q_k(w) = f(w_{k-1}) + \nabla f(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2\eta_k} \|w - w_{k-1}\|_2^2.$$

  upper bound of $f(w)$
- Locally a more accurate approximation of $f(x)$ is to use Hessian:

$$Q_k(w) = f(w_{k-1}) + \nabla f(w_{k-1})^T (w - w_{k-1}) + \frac{1}{2}(w - w_{k-1})^T H (w - w_{k-1}).$$

- BFGS; approximate H using first order gradients.
- LBFGS: use limited memory (store a few vectors) to approximate H
- Very effective for optimization of smooth objective functions.

Coordinate Descent (CD)

Let $f(w) = f([w_1, \ldots, w_d])$
Algorithm:

- for $j = 1, \ldots, d$
  - $w_j \leftarrow \arg\min_u f([w_1, \ldots, w_{j-1}, u, w_{j+1}, \ldots, w_d])$
- repeat until convergence

Idea: optimize one parameter at a time and fix others

Coordinate Descent (CD)

Let $f(w) = f([w_1, \ldots, w_d])$
Algorithm:

- for $j = 1, \ldots, d$
  - $w_j \leftarrow \arg\min_u f([w_1, \ldots, w_{j-1}, u, w_{j+1}, \ldots, w_d])$
- repeat until convergence

Idea: optimize one parameter at a time and fix others
Assumption:

- each one dimensional problem can be solved easily.
- each coordinate update for variable j is inexpensive compared to gradient descent.

Linear Regularization Problem

Consider regularized logistic regression:

$$w = \arg\min_w \left[\sum_{i=1}^{n} \ln(1 + \exp(-w^\top x_i y_i)) + \lambda \|w\|_1 \right]$$

or more generally the following problem with scaler functions f_i and h_j:

$$w = \arg\min_w \left[\sum_{i=1}^{n} f_i(w^T x_i) + \sum_{j=1}^{d} h_j(w_j) \right].$$

Iteration complexity:

- maintain $z_i = w^\top x_i$ for $i = 1, \ldots, n$
- coordinate j: update w_j and $\{z_i\}$ requires scanning a feature column
- one pass over $j = 1, \ldots, d$: one gradient descent step

Convergence

Practice:

- for suitable problems, coordinate descent works much better than gradient descent
- e.g., regularized logistic regression

Theory: incomplete

- current analysis either shows no improvements or improvements under very restricted scanerios
- Paul Tseng, Yurii Nesterov, ...

Convergence

Practice:

- for suitable problems, coordinate descent works much better than gradient descent
- e.g., regularized logistic regression

Theory: incomplete

- current analysis either shows no improvements or improvements under very restricted scanerios
- Paul Tseng, Yurii Nesterov, ...

It is still an open question to develop better theoretical understanding on when does coordinate descent performs better.

Convex Duality

- Given a convex function $f(w)$, we can define its conjugate or dual

$$f^*(\alpha) = \sup_w [w^T \alpha - f(w)].$$

The optimal w is $\alpha = \nabla f(w)$.

Convex Duality

- Given a convex function $f(w)$, we can define its conjugate or dual

$$f^*(\alpha) = \sup_w [w^T \alpha - f(w)].$$

  The optimal w is $\alpha = \nabla f(w)$.
- The dual of f^* is f:

$$f(w) = \sup_\alpha [w^T \alpha - f^*(\alpha)].$$

  The optimal α is $\nabla f^*(\alpha) = w$.

Convex Duality

- Given a convex function $f(w)$, we can define its conjugate or dual

$$f^*(\alpha) = \sup_w [w^T \alpha - f(w)].$$

  The optimal w is $\alpha = \nabla f(w)$.
- The dual of f^* is f:

$$f(w) = \sup_\alpha [w^T \alpha - f^*(\alpha)].$$

  The optimal α is $\nabla f^*(\alpha) = w$.
- We have the following property: for all w and α:

$$f(w) + f^*(\alpha) \geq w^T \alpha$$

  equality holds only at $w = \nabla f^*(\alpha)$: equivalent to $\alpha = \nabla f(w)$.

Dual of Linear Regularization Method

Primal optimization problem:

$$w_* = \arg\min_w P(w) \qquad P(w) := \sum_{i=1}^n f_i(w^T x_i) + \lambda g(w).$$

Dual optimization problem:

$$\alpha_* = \arg\max_\alpha D(\alpha) \qquad D(\alpha) = \sum_{i=1}^n -f_i^*(-\alpha_i) - \lambda g^*(\lambda^{-1}\sum_i \alpha_i x_i).$$

Strong duality:

- $P(w) \geq D(\alpha)$ for all w and α
- $P(w_*) = D(\alpha_*)$ with the relationship:

$$w_* = \nabla g^*\left(\lambda^{-1}\sum_{i=1}^n \alpha_{*,i} x_i\right) \qquad \alpha_{*\,i} = f_i'(w_*^T x_i).$$

Solve dual instead of primal problem.

Quick Justification of Strong Duality

$$
\begin{aligned}
P(w) - D(\alpha) = & \left[\sum_{i=1}^{n} f_i(w^T x_i) + \lambda g(w)\right] \\
& - \left[\sum_{i=1}^{n} -f_i^*(-\alpha_i) - \lambda g^*\left(\lambda^{-1}\sum_{i=1}^{n}\alpha_i x_i\right)\right] \\
= & \sum_{i=1}^{n}\left[f_i(w^T x_i) + f_i^*(\alpha_i) - \alpha_i w^T x_i\right] \\
& + \lambda\left[g(w) + g^*\left(\lambda^{-1}\sum_{i=1}^{n}\alpha_i x_i\right) - w^T\left(\lambda^{-1}\sum_{i=1}^{n}\alpha_i x_i\right)\right] \geq 0.
\end{aligned}
$$

Equality holds at $f_i'(w^T x_i) = \alpha_i$ and $w = \nabla g^*(\lambda^{-1}\sum_i \alpha_i x_i)$.

Quick Justification of Strong Duality

$$
\begin{aligned}
P(w) - D(\alpha) = & \left[\sum_{i=1}^{n} f_i(w^T x_i) + \lambda g(w)\right] \\
& - \left[\sum_{i=1}^{n} -f_i^*(-\alpha_i) - \lambda g^*\left(\lambda^{-1}\sum_{i=1}^{n}\alpha_i x_i\right)\right] \\
= & \sum_{i=1}^{n}\left[f_i(w^T x_i) + f_i^*(\alpha_i) - \alpha_i w^T x_i\right] \\
& + \lambda\left[g(w) + g^*\left(\lambda^{-1}\sum_{i=1}^{n}\alpha_i x_i\right) - w^T\left(\lambda^{-1}\sum_{i=1}^{n}\alpha_i x_i\right)\right] \geq 0.
\end{aligned}
$$

Equality holds at $f_i'(w^T x_i) = \alpha_i$ and $w = \nabla g^*(\lambda^{-1}\sum_i \alpha_i x_i)$.
Can check this gives the first order optimality conditions for w_* and α_*.

Example: Linear Support Vector Machine

- Primal formulation:

$$P(w) = \sum_{i=1}^{n} (1 - w^\top x_i y_i)_+ + 0.5\lambda \|w\|_2^2$$

  - $f_i(u) = (1 - u y_i)_+$
  - $g(w) = 0.5\|w\|_2^2$.
- Dual formulation:

$$D(\alpha) = \sum_{i=1}^{n} \alpha_i y_i + 0.5\lambda^{-1} \left\| \sum_{i=1}^{n} \alpha_i x_i y_i \right\|_2^2, \quad \alpha_i y_i \in [0, 1].$$

  - $-f_i^*(\alpha_i) = \alpha_i y_i$ with constraint $\alpha_i y_i \in [0, 1]$
  - $g^*(w) = 0.5\|w\|_2^2$

Dual Coordinate Descent

Dual optimization problem:

$$\alpha_* = \arg\max_{\alpha} D(\alpha) \qquad D(\alpha) = \sum_{i=1}^{n} -f_i^*(-\alpha_i) - \lambda g^*(\lambda^{-1}\sum_i \alpha_i x_i).$$

Apply coordinate descent on dual:

- maintain $w = \lambda^{-1}\sum_i \alpha_i x_i$
- for $i = 1, \ldots, n$, we update α_i one at a time while fixing the others

Computation: total computation of one pass over the data is comparable to one gradient descent.

Convergence

Previous analysis of the method only shows slow convergence.

Convergence

Previous analysis of the method only shows slow convergence.

Our new analysis (work in process with Shai Shalev-Schwartz):
To achieve accuracy ϵ

- for smooth loss (e.g. logistic), requires

$$O\left(\ln n + \frac{\ln(1/\epsilon)}{n}\right) \qquad \text{passes over data}$$

  - gradient descent: $O(\ln(1/\epsilon))$

Convergence

Previous analysis of the method only shows slow convergence.

Our new analysis (work in process with Shai Shalev-Schwartz):
To achieve accuracy ϵ

- for smooth loss (e.g. logistic), requires

$$O\left(\ln n + \frac{\ln(1/\epsilon)}{n}\right) \qquad \text{passes over data}$$

  - gradient descent: $O(\ln(1/\epsilon))$
- for nonsmooth loss (.e.g, SVM), requires

$$O\left(\ln n + \frac{1}{n\epsilon}\right) \qquad \text{passes over data}$$

  and convergence becomes geometric asymptotically
  - gradient descent: $O(1/\epsilon)$

References

- LBFGS: “On the limited memory BFGS method for large scale optimization”, Dong C. Liu and Jorge Nocedal, Mathematical Programming, 1989.
- Stephen Boyd and Lieven Vandenberghe: Convex Optimization Book (http://www.stanford.edu/ boyd/cvxbook/)
- Yurii Nesterov: proximal gradient and accelerated proximal gradient
  - Introductory Lectures on Convex Optimization: A Basic Course
  - Gradient methods for minimizing composite objective function
- Arkadi Nemirovski: optimization lecture notes http://www2.isye.gatech.edu/ nemirovs/