Optimization for Datascience

Proximal operator and proximal gradient methods

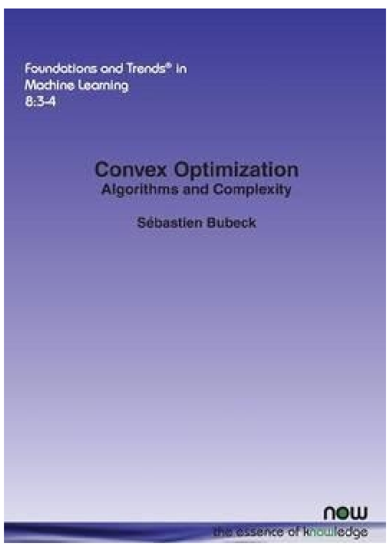
Lecturer: Robert M. Gower & Alexandre Gramfort

Tutorials: Quentin Bertrand, Nidham Gazagnadou

Master 2 Data Science, Institut Polytechnique de Paris (IPP)

References for todays class

Sébastien Bubeck (2015)
Convex Optimization: Algorithms and Complexity



Chapter 1 and Section 5.1

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

Optimization Sum of Terms

A Datum Function

$$f_i(w) := \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) &= \frac{1}{n}\sum_{i=1}^{n} \left(\ell\left(h_w(x^i), y^i\right) + \lambda R(w)\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} f_i(w)\end{aligned}$$

Finite Sum Training Problem

$$\min_{w \in \mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

The Training Problem

Solving the *training problem*:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Reference method: Gradient descent

$$\nabla \left(\frac{1}{n} \sum_{i=1}^{n} f_i(w) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w)$$

Gradient Descent Algorithm

Set $w^1 = 0$, choose $\alpha > 0$.
for $t = 1, 2, 3, \ldots, T$
$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$$
Output w^{T+1}

Convergence GD I

Theorem

Let f be convex and L-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{T-1} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

Convergence GD I

Theorem

Let f be convex and L-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{T-1} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

Is f always differentiable?

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

Convergence GD I

Not true for many problems

Theorem

Let f be convex and L-smooth.

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{T-1} = O\left(\frac{1}{T}\right).$$

Where

$$w^{t+1} = w^t - \frac{1}{L}\nabla f(w^t)$$

Is f always differentiable?

$$\Rightarrow \text{for } \frac{f(w^T) - f(w^*)}{||w^1 - w^*||_2^2} \leq \epsilon \text{ we need } T \geq \frac{2L}{\epsilon} = O\left(\frac{1}{\epsilon}\right)$$

Change notation: Keep loss and regularizor separate

Loss function

$$L(w) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right)$$

The Training problem

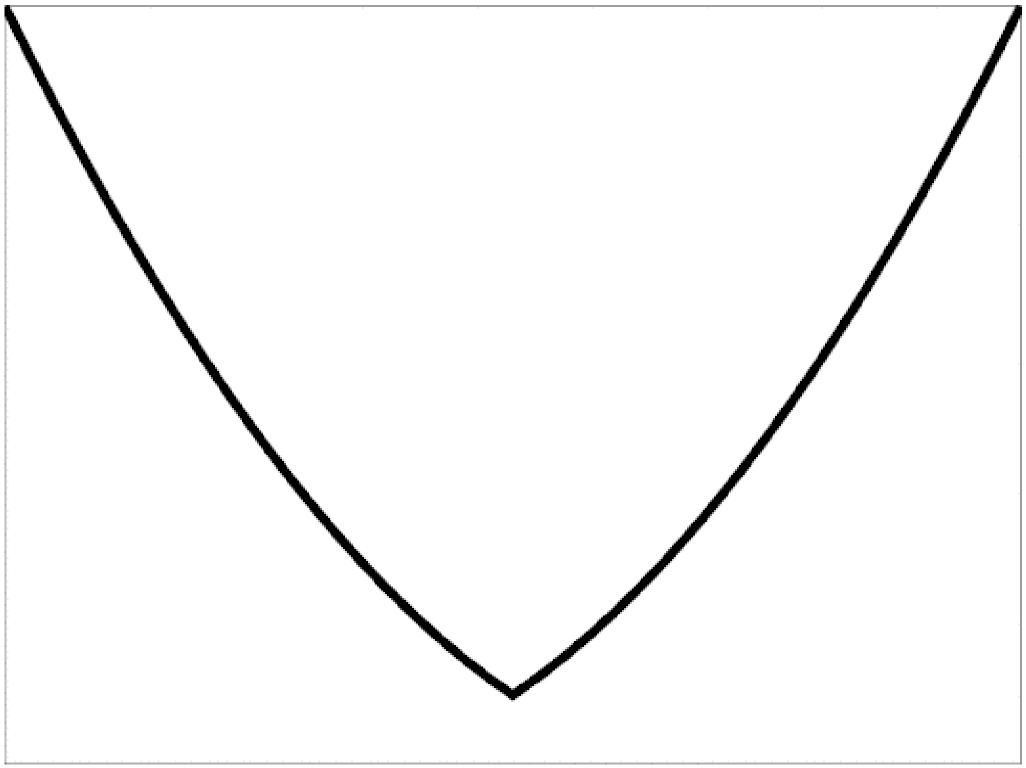
$$\min_w L(w) + \lambda R(w)$$

If L or R is not differentiable → $L+R$ is not differentiable

If L or R is not smooth → $L+R$ is not smooth

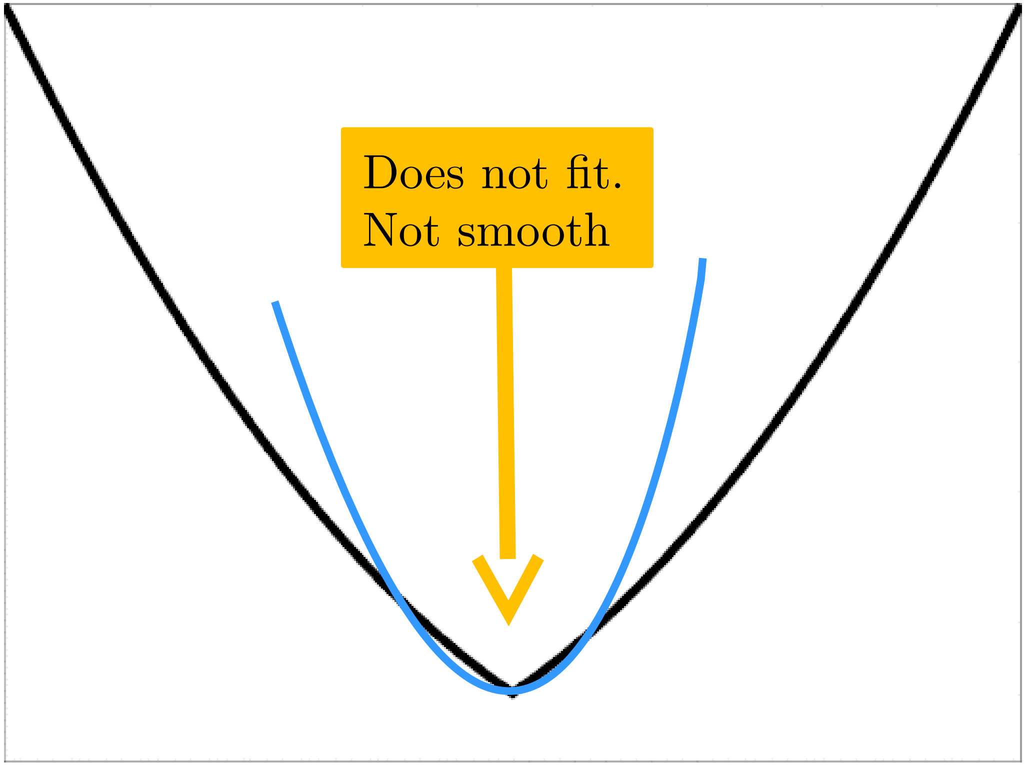
Non-smooth Example

$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

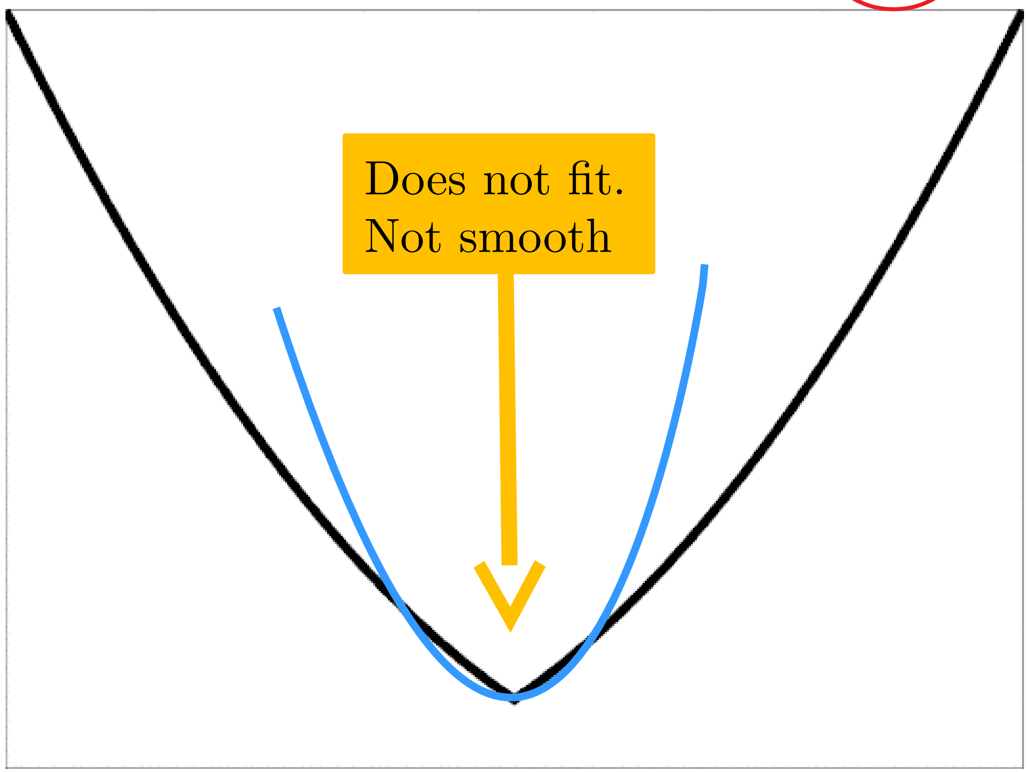
$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

The problem

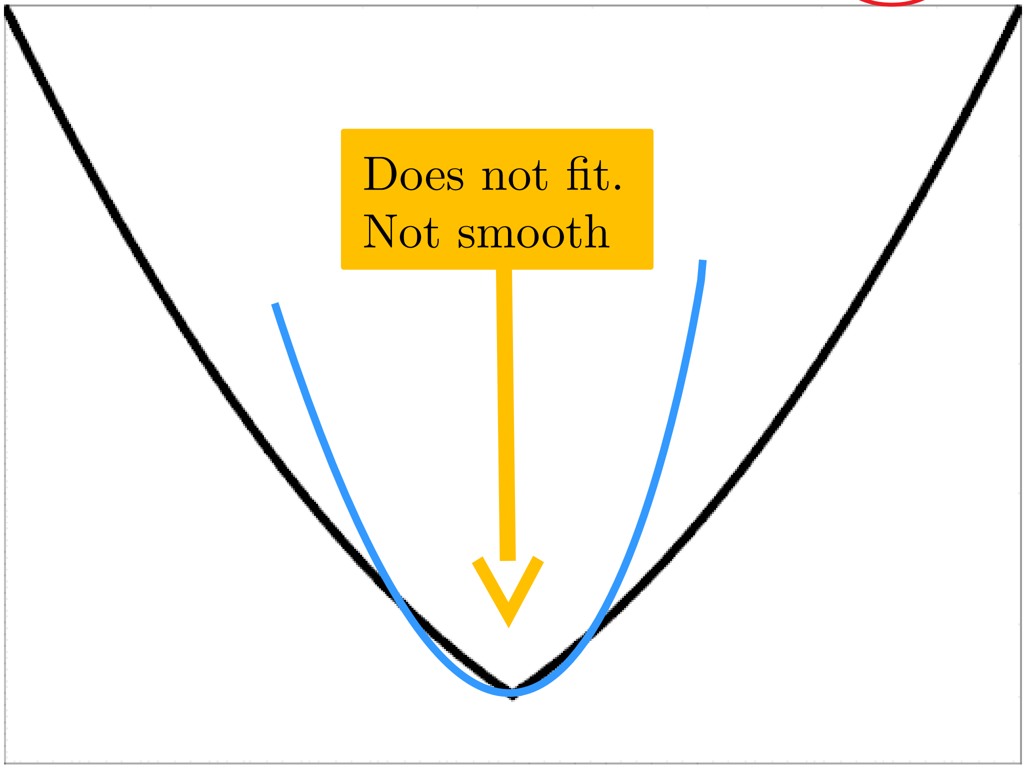
$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Non-smooth Example

The problem

$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$

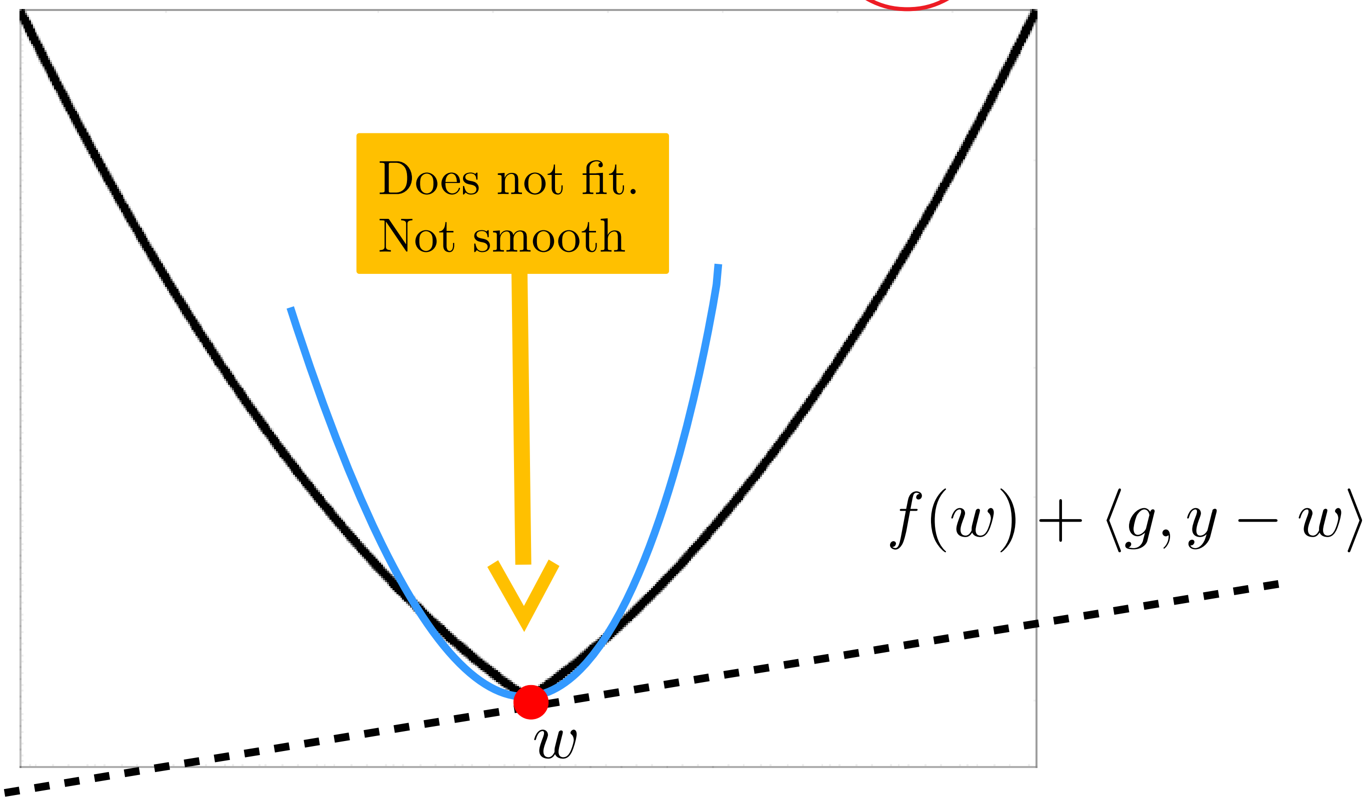


Need more tools

Non-smooth Example

The problem

$$L(w) + R(w) = \frac{1}{2}||w||_2^2 + ||w||_1$$



Need more tools

Assumptions for this class

The Training problem

$$\min_{w} L(w) + \lambda R(w)$$

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and "easy to optimize"

What does this mean?

$$\mathrm{prox}_{\gamma R}(y) := \arg\min_{w} \frac{1}{2}||w - y||_2^2 + \gamma R(w)$$

Assume this is easy to solve

Examples

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2n} \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 + \lambda ||w||_1$$

Low Rank Matrix Recovery

$$\min_{W \in \mathbf{R}^{d \times d}} \frac{1}{n} \sum_{i=1}^{n} ||AW - Y||_F^2 + \lambda ||W||_*$$

Not smooth, but prox is easy

SVM with soft margin

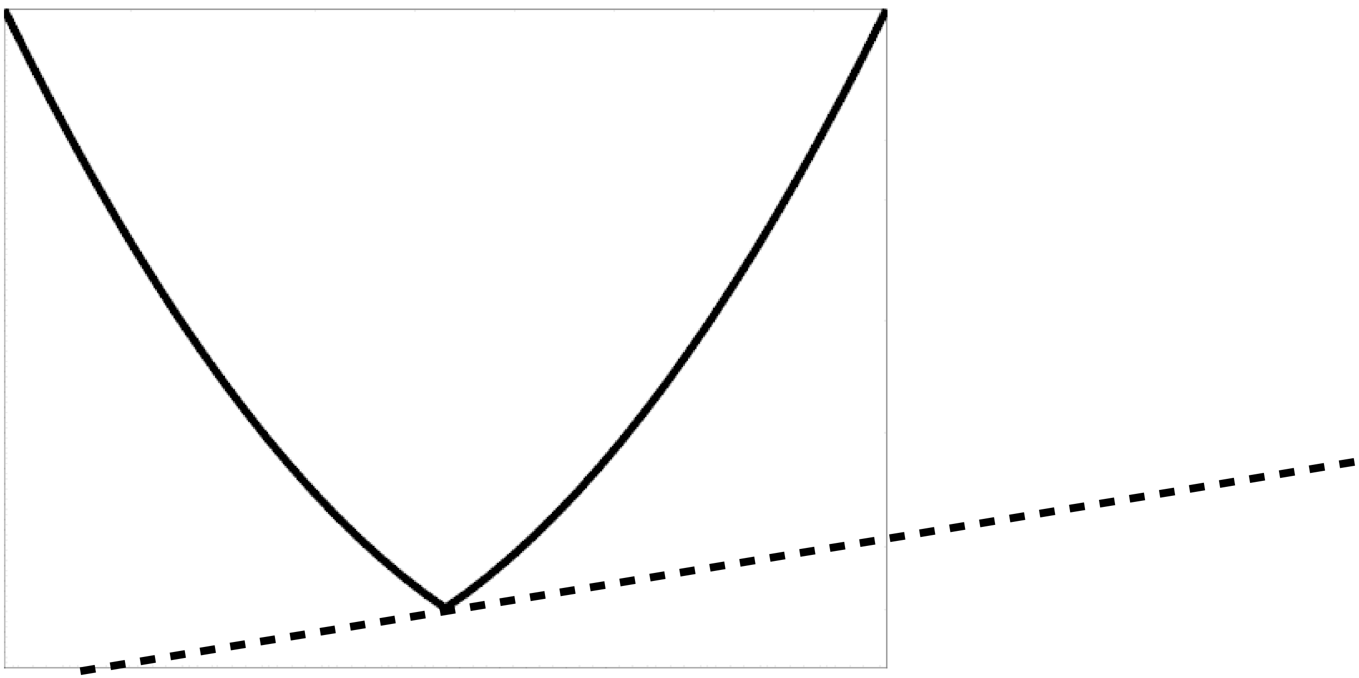
$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \max\{0, 1 - y^i \langle w, a^i \rangle\} + \lambda ||w||_2^2$$

Not smooth

Convexity: Subgradient

Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n \ : \ f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \mathrm{dom}(f)\}$$

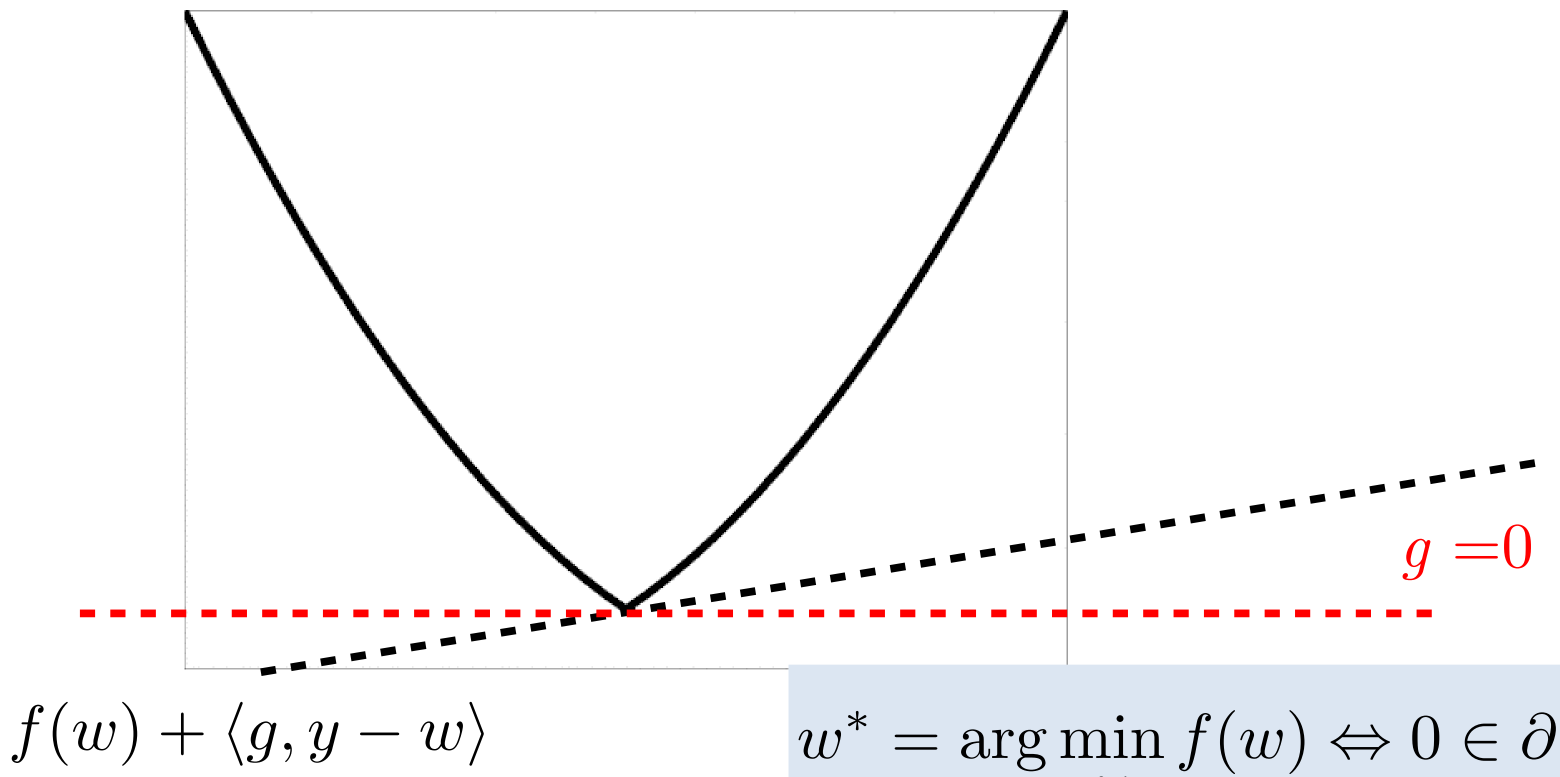


$f(w) + \langle g, y - w \rangle$

Convexity: Subgradient

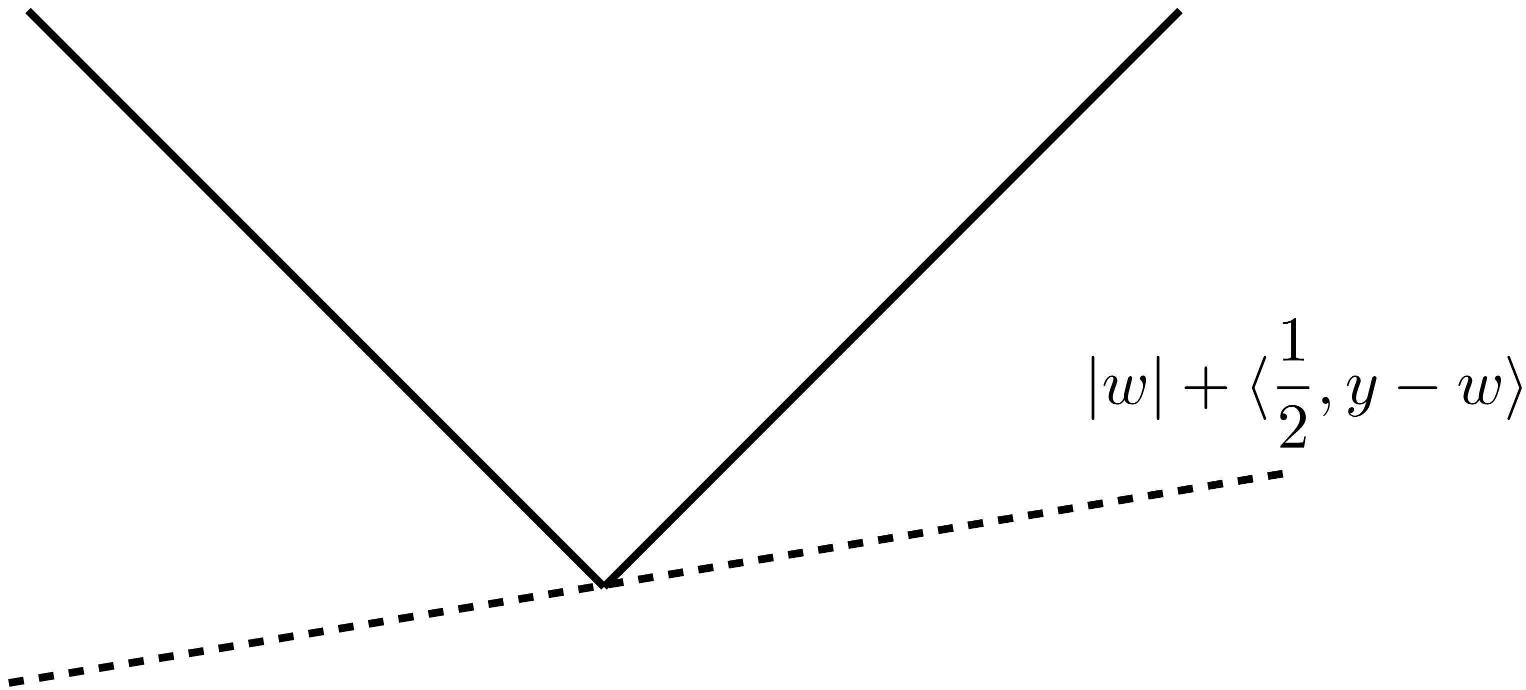
Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ be convex

$$\partial f(w) := \{g \in \mathbb{R}^n \, : \, f(y) \geq f(w) + \langle g, y - w \rangle, \forall y \in \mathrm{dom}(f)\}$$



$$w^* = \arg\min_w f(w) \Leftrightarrow 0 \in \partial f(w^*)$$

Examples: L1 norm



$$\partial |w| = \begin{cases} -1 & \text{if } w < 0 \\ [-1, 1] & \text{if } w = 0 \\ 1 & \text{if } w > 0 \end{cases}$$

$$\partial ||w||_1 = (\partial |w_1|, \ldots, \partial |w_d|)$$

Optimality conditions

The Training problem

$$w^* = \arg \min_{w \in \mathbf{R}^d} L(w) + \lambda R(w)$$

$$0 \quad \in \quad \partial \left(L(w^*) + \lambda R(w^*)\right) = \nabla L(w^*) + \lambda \partial R(w^*)$$

$$\Downarrow$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*)$$

Working example: Lasso

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2n} ||Aw - y||_2^2 + \lambda ||w||_1$$

$$A = [a^1, \ldots, a^n]^\top \Rightarrow \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 = ||Aw - y||_2^2$$

$$-\nabla L(w^*) \in \partial R(w^*) \quad \Rightarrow \quad -\frac{1}{n} A^\top (Aw^* - y) \in \partial ||w^*||_1$$

Difficult inclusion, do iteratively.

Proximal method I

Using $\mathcal{L}$–smoothness of L :

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

Proximal method I

Using $\mathcal{L}$–smoothness of L :

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

But what about *R(w)*? Adding on $+\lambda R(w)$ to upper bound:

Proximal method I

Using $\mathcal{L}$–smoothness of L :

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2}||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

But what about $R(w)$? Adding on $+\lambda R(w)$ to upper bound:

$$L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2}||w - y||^2 + \lambda R(w)$$

Proximal method I

Using $\mathcal{L}$–smoothness of L :

$$L(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2, \quad \forall w, y \in \mathbb{R}^d$$

The w that minimizes the upper bound gives gradient descent

$$w = y - \frac{1}{\mathcal{L}} \nabla L(y)$$

But what about $R(w)$? Adding on $+\lambda R(w)$ to upper bound:

$$L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w)$$

Can we minimize the right-hand side?

Proximal method: iteratively minimizes an upper bound

Minimizing the right-hand side of

$$L(w) + \lambda R(w) \le L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2}||w - y||^2 + \lambda R(w)$$

$$\begin{aligned} &\arg\min_w L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2}||w - y||^2 + \lambda R(w) \\ &= \arg\min_w \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2}||w - y||^2 + \lambda R(w) \\ &= \arg\min_w \frac{1}{2}||w - (y - \frac{1}{\mathcal{L}}\nabla L(y))||^2 + \frac{\lambda}{\mathcal{L}} R(w) \\ &=: \text{prox}_{\frac{\lambda}{\mathcal{L}} R}(y - \frac{1}{\mathcal{L}}\nabla L(y))) \end{aligned}$$

$$\text{prox}_{\frac{\lambda}{\mathcal{L}} R}(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + \frac{\lambda}{\mathcal{L}} R(w)$$

Proximal method: minimizes an upperbound viewpoint

Set $y = w^t$ and minimize the right-hand side in w

$$L(w) + \lambda R(w) \leq L(w^t) + \langle \nabla L(w^t), w - w^t \rangle + \frac{\mathcal{L}}{2} ||w - w^t||^2 + \lambda R(w)$$

$$\arg\min_w L(w^t) + \langle \nabla L(w^t), w - w^t \rangle + \frac{\mathcal{L}}{2} ||w - w^t||^2 + \lambda R(w)$$

$$=: \text{prox}_{\frac{\lambda}{\mathcal{L}} R}(w^t - \frac{1}{\mathcal{L}} \nabla L(w^t)))$$

This suggests an iterative method

$$w^{t+1} = \text{prox}_{\frac{\lambda}{\mathcal{L}} R}(w^t - \frac{1}{\mathcal{L}} \nabla L(w^t)))$$

Proximal method: minimizes an upperbound viewpoint

Set $y = w^t$ and minimize the right-hand side in w

$$L(w) + \lambda R(w) \leq L(w^t) + \langle \nabla L(w^t), w - w^t \rangle + \frac{\mathcal{L}}{2} ||w - w^t||^2 + \lambda R(w)$$

$$\arg\min_w L(w^t) + \langle \nabla L(w^t), w - w^t \rangle + \frac{\mathcal{L}}{2} ||w - w^t||^2 + \lambda R(w)$$

$$=: \mathrm{prox}_{\frac{\lambda}{\mathcal{L}} R}(w^t - \frac{1}{\mathcal{L}} \nabla L(w^t))$$

This suggests an iterative method

What is this prox operator?

$$w^{t+1} = \mathrm{prox}_{\frac{\lambda}{\mathcal{L}} R}(w^t - \frac{1}{\mathcal{L}} \nabla L(w^t))$$

Gradient Descent using proximal map

$$\text{prox}_{\gamma R}(y) := \arg\min_{w} \frac{1}{2}||w - y||_2^2 + \gamma R(w)$$

EXE: **Let**

$$R(w) \;=\; f(y) + \langle \nabla f(y), w - y \rangle$$

Show that

$$\text{prox}_{\gamma R}(y) = y - \gamma \nabla f(y)$$

A gradient step is also a proximal step

Proximal Operator: Well defined inclusion

Let $f(x)$ be a convex function. The proximal operator is

$$\text{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Let $w_v = \text{prox}_f(v)$.

EXE: Is this Proximal operator well defined? Is it even a function?

Proximal Operator: Well defined inclusion

Let $f(x)$ be a convex function. The proximal operator is

$$\mathrm{prox}_f(v) := \arg\min_w \frac{1}{2}||w-v||_2^2 + f(w)$$

Let $w_v = \mathrm{prox}_f(v)$. Using optimality conditions

$$0 \in \partial\left(\tfrac{1}{2}||w_v - v||_2^2 + f(w)\right) = w_v - v + \partial f(w_v)$$

EXE: Is this Proximal operator well defined? Is it even a function?

Proximal Operator: Well defined inclusion

Let $f(x)$ be a convex function. The proximal operator is

$$\text{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Let $w_v = \text{prox}_f(v)$. Using optimality conditions

$$0 \in \partial\left(\tfrac{1}{2}||w_v - v||_2^2 + f(w)\right) = w_v - v + \partial f(w_v)$$

Rearranging

$$\text{prox}_f(v) = w_v \in v - \partial f(w_v)$$

EXE: Is this Proximal operator well defined? Is it even a function?

Proximal Method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_{w} L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*)$$

Proximal Method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_{w} L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*) \quad \Longleftrightarrow \quad w^* + \gamma \nabla L(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

Proximal Method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_w L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla L(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla L(w^*)) - (\lambda\gamma)\partial R(w^*)$$

Proximal Method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_{w} L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla L(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla L(w^*)) - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* = \mathrm{prox}_{\lambda\gamma R}(w^* - \gamma \nabla L(w^*))$$

$$\mathrm{prox}_f(v) = w_v \in v - \partial f(w_v)$$

Proximal Method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_{w} L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla L(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla L(w^*)) - (\lambda\gamma)\partial R(w^*)$$

$\text{prox}_f(v) = w_v \in v - \partial f(w_v)$

$$\iff w^* = \text{prox}_{\lambda\gamma R}\left(w^* - \gamma \nabla L(w^*)\right)$$

Optimal is a fixed point

$$\Rightarrow w^{k+1} = \text{prox}_{\lambda\gamma R}\left(w^k - \gamma \nabla L(w^k)\right)$$

Proximal Method: A fixed point viewpoint

The Training problem

$$w^* \in \arg\min_{w} L(w) + \lambda R(w)$$

$$-\nabla L(w^*) \in \lambda \partial R(w^*) \iff w^* + \gamma \nabla L(w^*) \in w^* - (\lambda\gamma)\partial R(w^*)$$

$$\iff w^* \in (w^* - \gamma \nabla L(w^*)) - (\lambda\gamma)\partial R(w^*)$$

$\text{prox}_f(v) = w_v \in v - \partial f(w_v)$

$$\iff w^* = \text{prox}_{\lambda\gamma R}\left(w^* - \gamma \nabla L(w^*)\right)$$

Optimal is a fixed point $\Rightarrow$

$$w^{k+1} = \text{prox}_{\lambda\gamma R}\left(w^k - \gamma \nabla L(w^k)\right)$$

Upperbound viewpoint $\Rightarrow$

$$w^{t+1} = \text{prox}_{\frac{\lambda}{\mathcal{L}} R}(w^t - \frac{1}{\mathcal{L}} \nabla L(w^t)))$$

Proximal Operator: Properties

$$\text{prox}_f(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + f(w)$$

Exe:

1) If $f(w) = \sum_{i=1}^{d} f_i(w_i)$ then $\text{prox}_f(v) = (\text{prox}_{f_1}(v_1), \ldots, \text{prox}_{f_d}(v_d))$

2) If $f(w) = I_C(w) := \begin{cases} 0 & \text{if } w \in C \\ \infty & \text{if } w \notin C \end{cases}$ where C closed and convex

then $\text{prox}_f(v) = \text{proj}_C(v)$

3) If $f(w) = \langle b, w \rangle + c$ then $\text{prox}_f(v) = v - b$

4) If $f(w) = \frac{\lambda}{2} w^\top A w + \langle b, w \rangle$ where $A \succeq 0,\ A = A^\top, \lambda \geq 0$ then

$$\text{prox}_f(v) = (I + \lambda A)^{-1}(v - b)$$

Proximal Operator: Soft thresholding

$$\text{prox}_{\lambda||w||_1}(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + \lambda||w||_1$$

Exe:

1) Let $\alpha \in \mathbf{R}$. If $\alpha^* = \arg\min_\alpha \frac{1}{2}(\alpha - v)^2 + \lambda|\alpha|$ then

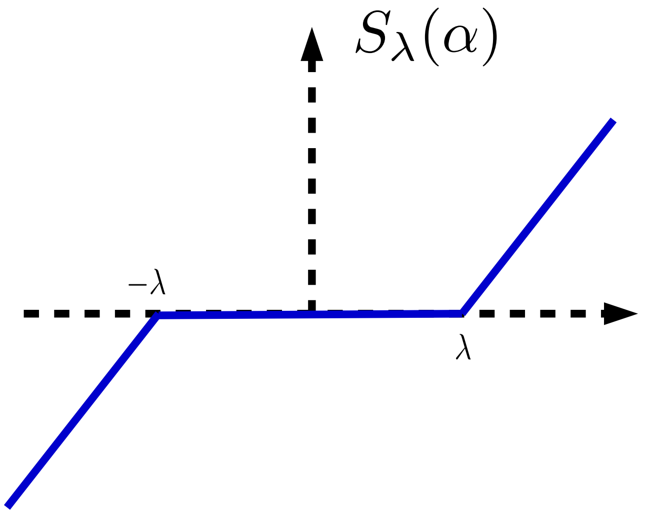
$$\alpha^* \in v - \lambda\partial|\alpha^*| \qquad (I)$$

2) If $\lambda < v$ show (I) gives $\alpha^* = v - \lambda$

3) If $v < -\lambda$ show (I) gives $\alpha^* = v + \lambda$

4) Show that

$$\text{prox}_{\lambda|\alpha|}(v) = \begin{cases} v - \lambda & \text{if } \lambda < v \\ 0 & \text{if } -\lambda \leq v \leq \lambda \\ v + \lambda & \text{if } v < -\lambda. \end{cases}$$



Proximal Operator: Singular value thresholding

$$S_\lambda(v) := \arg\min_w \frac{1}{2}||w - v||_2^2 + \lambda||w||_1$$

Similarly, the prox of the nuclear norm for matrices:

$$U\text{diag}(S_\lambda(\text{diag}(\sigma(A))))V^\top := \arg\min_{W \in \mathbf{R}^{d \times d}} \frac{1}{2}||W - A||_F^2 + \lambda||W||_*$$

where $A = U\text{diag}(\sigma(A))V^\top$ is a SVD decomposition,

and $\|W\|_* = \text{trace}(\sqrt{W^\top W}) = \sum_i \sigma_i(W)$ is the nuclear norm

Proximal method: iteratively minimizes an upper bound

Minimizing the right-hand side of

$$L(w) + \lambda R(w) \leq L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w)$$

$$\arg\min_w L(y) + \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w)$$

$$= \arg\min_w \langle \nabla L(y), w - y \rangle + \frac{\mathcal{L}}{2} ||w - y||^2 + \lambda R(w)$$

$$= \arg\min_w \frac{1}{2} ||w - (y - \frac{1}{\mathcal{L}} \nabla L(y)||^2 + \frac{\lambda}{\mathcal{L}} R(w)$$

$$= \text{prox}_{\frac{\lambda}{\mathcal{L}} R} \left(y - \frac{1}{\mathcal{L}} \nabla L(y) \right)$$

Make iterative method based on this upper bound minimization

The Proximal Gradient Method

Solving the *training problem*:

$$\min_w L(w) + \lambda R(w)$$

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and prox friendly

Proximal Gradient Descent

Set $w^1 = 0$.
for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = \mathrm{prox}_{\lambda R/\mathcal{L}}\left(w^t - \frac{1}{\mathcal{L}}\nabla L(w^t)\right)$$

Output w^{T+1}

Example of prox gradient: Iterative Soft Thresholding Algorithm (ISTA)

Lasso

$$\min_{w \in \mathbf{R}^d} \frac{1}{2n} ||Aw - y||_2^2 + \lambda ||w||_1$$

$$A = [a^1, \ldots, a^n]^\top \Rightarrow \sum_{i=1}^{n} (y^i - \langle w, a^i \rangle)^2 = ||Aw - y||_2^2$$

ISTA:

$$w^{t+1} = \text{prox}_{\lambda ||w||_1 / \mathcal{L}} \left(w^t - \frac{1}{n\mathcal{L}} A^\top (Aw^t - y) \right)$$

$$\mathcal{L} = \frac{\sigma_{\max}(A)^2}{n}$$

$$= S_{\lambda/\mathcal{L}} \left(w^t - \frac{1}{\sigma_{\max}(A)^2} A^\top (Aw^t - y) \right)$$

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

Convergence of Prox-GD

Theorem (Beck Teboulle 2009)

Let $f(w) = L(w) + \lambda R(w)$ where

$$L(w) \text{ is differentiable, } \mathcal{L}\text{–smooth and convex}$$

$$R(w) \text{ is convex and prox friendly}$$

Then

$$f(w^T) - f(w^*) \leq \frac{L||w^1 - w^*||_2^2}{2T} = O\left(\frac{1}{T}\right).$$

where

$$w^{t+1} = w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}}\left(w^t - \frac{1}{\mathcal{L}}\nabla L(w^t)\right)$$

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

Convergence of Prox-GD

Theorem (Beck Teboulle 2009)

Let $f(w) = L(w) + \lambda R(w)$ where

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and prox friendly

Can we do better?

Then

$$f(w^T) - f(w^*) \leq \frac{L||w^1 - w^*||_2^2}{2T} = O\left(\frac{1}{T}\right).$$

where

$$w^{t+1} = w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}}\left(w^t - \frac{1}{\mathcal{L}}\nabla L(w^t)\right)$$

PDF Adobe
Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm
for Linear Inverse Problems.

The FISTA Method

Solving the *training problem*:

$$\min_w L(w) + \lambda R(w)$$

The FISTA Algorithm

Set $w^1 = 0 = z^1, \beta^1 = 1$

for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}}\left(z^t - \frac{1}{\mathcal{L}}\nabla L(z^t)\right)$$

$$\beta^{t+1} = \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2}$$

$$z^{t+1} = w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}}(w^{t+1} - w^t)$$

Output w^{T+1}

The FISTA Method

Solving the *training problem*:

$$\min_w L(w) + \lambda R(w)$$

The FISTA Algorithm

Set $w^1 = 0 = z^1, \beta^1 = 1$

for $t = 1, 2, 3, \ldots, T$

$$w^{t+1} = \text{prox}_{\lambda R/\mathcal{L}}\left(z^t - \frac{1}{\mathcal{L}}\nabla L(z^t)\right)$$

$$\beta^{t+1} = \frac{1 + \sqrt{1 + 4(\beta^t)^2}}{2}$$

$$z^{t+1} = w^{t+1} + \frac{\beta^t - 1}{\beta^{t+1}}(w^{t+1} - w^t)$$

Output w^{T+1}

Weird, but it works

Convergence of FISTA

Theorem (Beck Teboulle 2009)

Let $f(w) = L(w) + \lambda R(w)$ where

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and prox friendly

Then

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Where w^t are given by the FISTA algorithm

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

Convergence of FISTA

Theorem (Beck Teboulle 2009)

Let $f(w) = L(w) + \lambda R(w)$ where

$L(w)$ is differentiable, $\mathcal{L}$–smooth and convex

$R(w)$ is convex and prox friendly

Is this as good as it gets?

Then

$$f(w^T) - f(w^*) \leq \frac{2L||w^1 - w^*||_2^2}{(T+1)^2} = O\left(\frac{1}{T^2}\right).$$

Where w^t are given by the FISTA algorithm

Amir Beck and Marc Teboulle (2009), SIAM J. IMAGING SCIENCES,
A Fast Iterative Shrinkage-Thresholding Algorithm for Linear Inverse Problems.

Lab Session 30.09

Room **C129** and **C130**

Bring your laptop

Please install:
Python, matplotlib, scipy and numpy

Lab Session 30.09

Room **C129** and **C130**

Bring your laptop

Please install:

Python, matplotlib, scipy and numpy

Introduction to Stochastic Gradient Descent

Recap

Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) =: f(w)$$

$L(w)$

General methods

$\min f(w)$

- Gradient Descent

Two parts

$\min L(w) + \lambda R(w)$

- Proximal gradient (ISTA)
- Fast proximal gradient (FISTA)

Optimization Sum of Terms

A Datum Function

$$f_i(w) := \ell\left(h_w(x^i), y^i\right) + \lambda R(w)$$

$$\frac{1}{n}\sum_{i=1}^{n} \ell\left(h_w(x^i), y^i\right) + \lambda R(w) = \frac{1}{n}\sum_{i=1}^{n}\left(\ell\left(h_w(x^i), y^i\right) + \lambda R(w)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} f_i(w)$$

Finite Sum Training Problem

$$\min_{w \in \mathbf{R}^d} \frac{1}{n}\sum_{i=1}^{n} f_i(w) =: f(w)$$

Can we use this sum structure?

The Training Problem

Solving the *training problem*: $$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Reference method: Gradient descent

$$\nabla \left(\frac{1}{n} \sum_{i=1}^{n} f_i(w) \right) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(w)$$

Gradient Descent Algorithm

Set $w^0 = 0$, choose $\alpha > 0$.

for $t = 0, 1, 2, \ldots, T-1$

$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$$

Output w^T

The Training Problem

Solving the *training problem*:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} f_i(w)$$

Problem with Gradient Descent:
Each iteration requires computing a gradient $\nabla f_i(w)$ for each data point. One gradient for each cat on the internet!

Gradient Descent Algorithm

Set $w^0 = 0$, choose $\alpha > 0$.
for $t = 0, 1, 2, \ldots, T$
$$w^{t+1} = w^t - \frac{\alpha}{n} \sum_{i=1}^{n} \nabla f_i(w^t)$$
Output w^T

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Unbiased Estimate

Let j be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Unbiased Estimate

Let j be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

Use $\nabla f_j(w) \approx \nabla f(w)$

Stochastic Gradient Descent

Is it possible to design a method that uses only the gradient of a **single** data function $f_i(w)$ at each iteration?

Unbiased Estimate

Let j be a random index sampled from $\{1, \ldots, n\}$ selected uniformly at random. Then

$$\mathbb{E}_j[\nabla f_j(w)] = \frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = \nabla f(w)$$

Use $\nabla f_j(w) \approx \nabla f(w)$

EXE: Let $\sum_{i=1}^{n} p_i = 1$ and $j \sim p_j$. Show $\mathbb{E}[\nabla f_j(w)/(np_j)] = \nabla f(w)$

Stochastic Gradient Descent

SGD 0.0 Constant stepsize

Set $w^0 = 0$, choose $\alpha > 0$

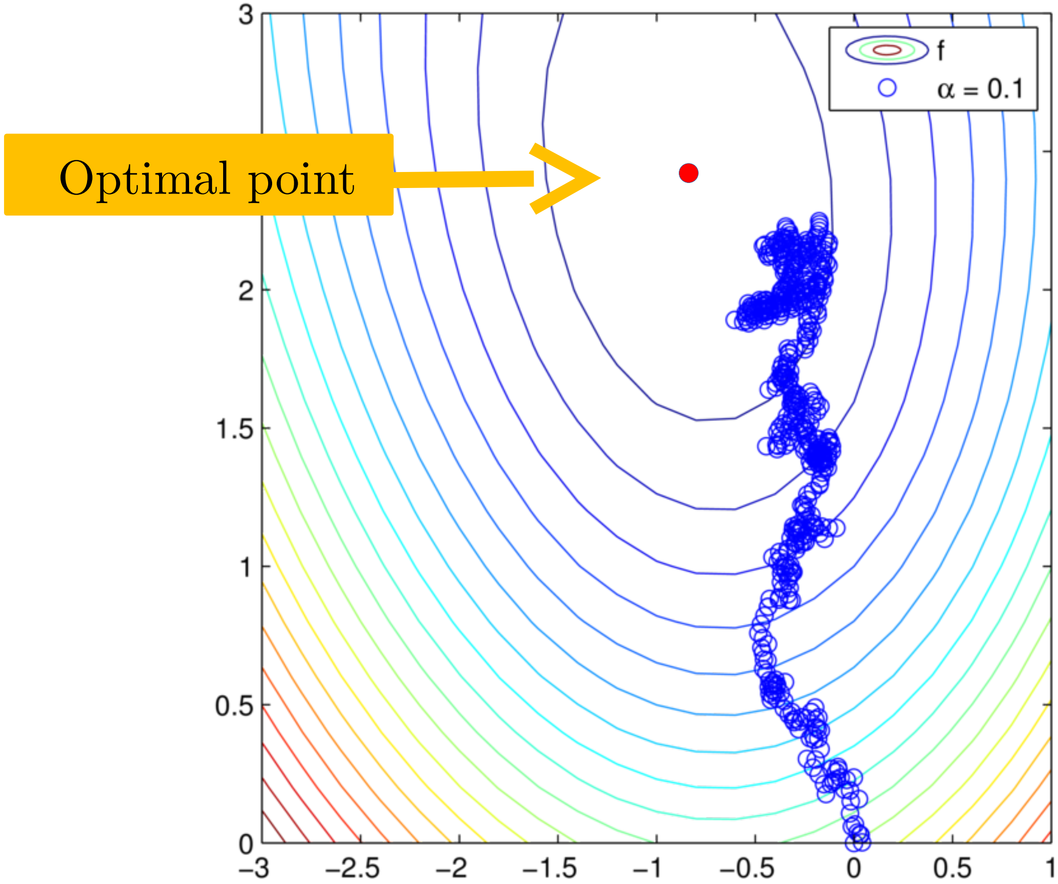
for $t = 0, 1, 2, \ldots, T-1$

  sample $j \in \{1, \ldots, n\}$

  $w^{t+1} = w^t - \alpha \nabla f_j(w^t)$

Output w^T

Stochastic Gradient Descent



Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2} ||y - w||_2^2, \quad \forall w, y$$

$$y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2}||y - w||_2^2, \quad \forall w, y$$

$$\Downarrow \; y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2}||y - w||_2^2, \quad \forall w, y$$

$$y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Expected Bounded Stochastic Gradients

$$\mathbb{E}_j[||\nabla f_j(w^t)||_2^2] \leq B^2, \text{ for all iterates } w^t \text{ of SGD}$$

Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2}||y - w||_2^2, \quad \forall w, y$$

$$\Downarrow \quad y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Expected Bounded Stochastic Gradients

$$\mathbb{E}_j[||\nabla f_j(w^t)||_2^2] \leq B^2, \text{ for all iterates } w^t \text{ of SGD}$$

Complexity / Convergence

Theorem

If $0 < \alpha \leq \frac{1}{\lambda}$ then the iterates of the SGD 0.0 method satisfy

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1 - \alpha\lambda)^t ||w^0 - w^*||_2^2 + \frac{\alpha}{\lambda} B^2$$

Complexity / Convergence

Theorem

If $0 < \alpha \leq \frac{1}{\lambda}$ then the iterates of the SGD 0.0 method satisfy

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1 - \alpha\lambda)^t ||w^0 - w^*||_2^2 + \frac{\alpha}{\lambda} B^2$$

Shows that $\alpha \approx \frac{1}{\lambda}$

Complexity / Convergence

Theorem

If $0 < \alpha \leq \frac{1}{\lambda}$ then the iterates of the SGD 0.0 method satisfy

$$\mathbb{E}\left[||w^t - w^*||_2^2\right] \leq (1 - \alpha\lambda)^t ||w^0 - w^*||_2^2 + \frac{\alpha}{\lambda} B^2$$

Shows that $\alpha \approx \frac{1}{\lambda}$

Shows that $\alpha \approx 0$

Proof:

$$||w^{t+1} - w^*||_2^2 = ||w^t - w^* - \alpha \nabla f_j(w^t)||_2^2$$

$$= ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f_j(w^t), w^t - w^* \rangle + \alpha^2 ||\nabla f_j(w^t)||_2^2.$$

Taking expectation with respect to j

Unbiased estimator

$$\mathbb{E}_j \left[||w^{t+1} - w^*||_2^2 \right] = ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 \mathbb{E}_j \left[||\nabla f_j(w^t)||_2^2 \right]$$

$$\leq ||w^t - w^*||_2^2 - 2\alpha \langle \nabla f(w^t), w^t - w^* \rangle + \alpha^2 B^2$$

Strong conv.

$$\leq (1 - \alpha\lambda)||w^t - w^*||_2^2 + \alpha^2 B^2$$

Bounded Stoch grad

Taking total expectation

$$\mathbb{E} \left[||w^{t+1} - w^*||_2^2 \right] \leq (1 - \alpha\lambda)\mathbb{E} \left[||w^t - w^*||_2^2 \right] + \alpha^2 B^2$$

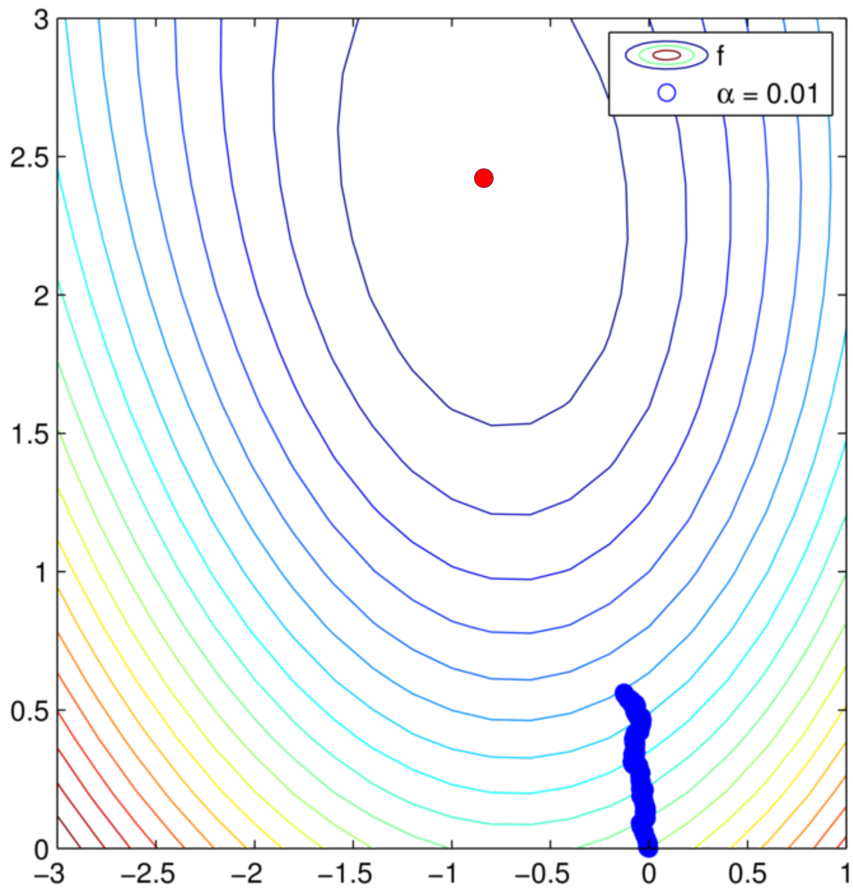
$$= (1 - \alpha\lambda)^{t+1} ||w^0 - w^*||_2^2 + \sum_{i=0}^t (1 - \alpha\lambda)^i \alpha^2 B^2$$

Using the geometric series sum $\displaystyle\sum_{i=0}^{t} (1 - \alpha\lambda)^i = \frac{1 - (1 - \alpha\lambda)^{t+1}}{\alpha\lambda} \leq \frac{1}{\alpha\lambda}$

$$\mathbb{E} \left[||w^{t+1} - w^*||_2^2 \right] \leq (1 - \alpha\lambda)^{t+1} ||w^0 - w^*||_2^2 + \frac{\alpha}{\lambda} B^2$$

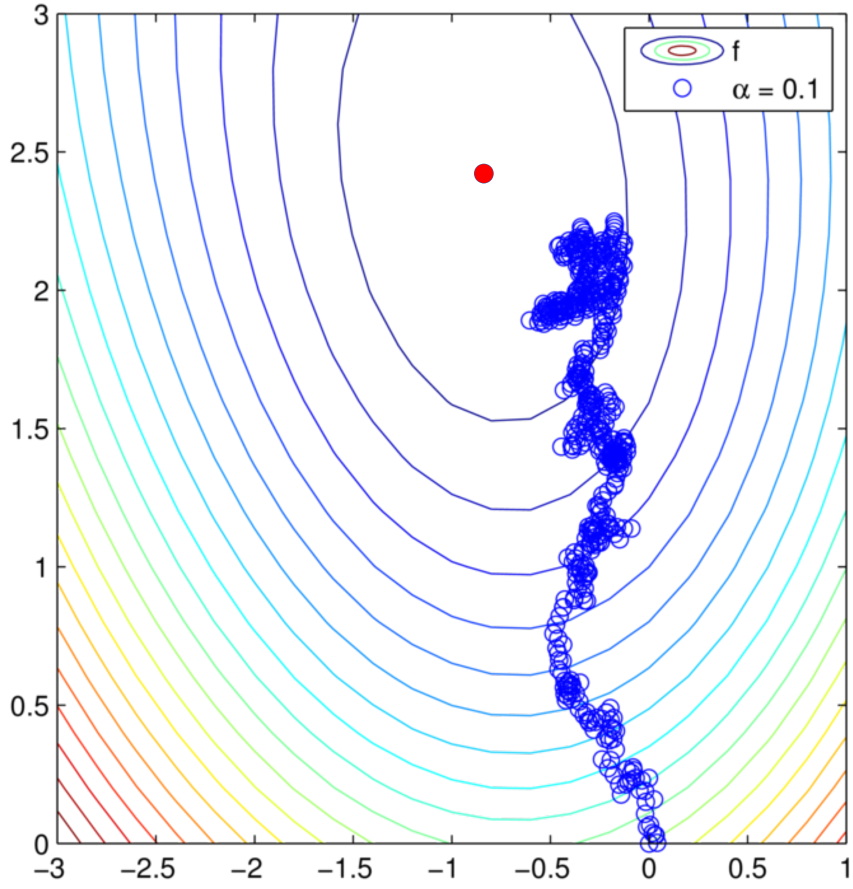
Stochastic Gradient Descent

α =0.01



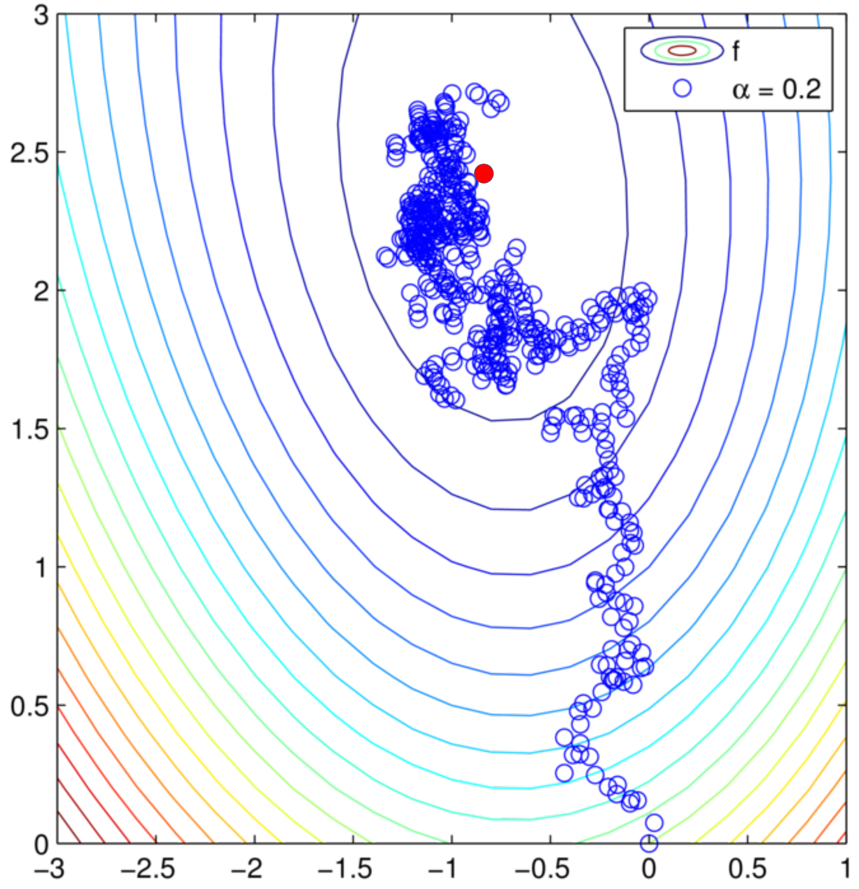
Stochastic Gradient Descent

α =0.1



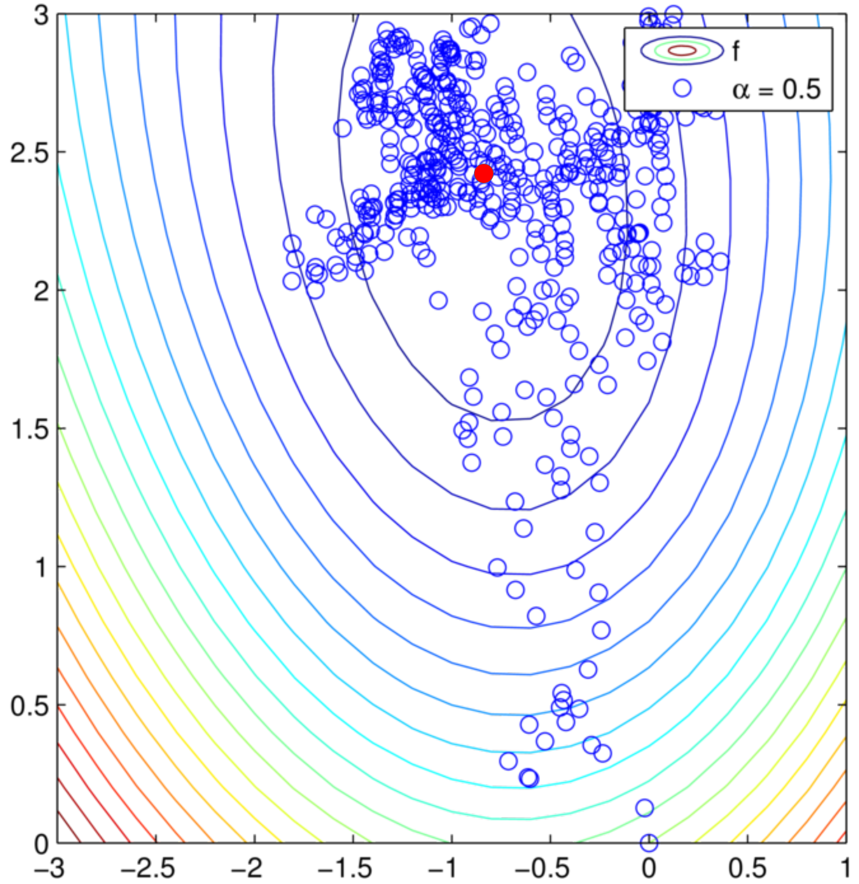
Stochastic Gradient Descent

α =0.2



Stochastic Gradient Descent

α =0.5



Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2}||y - w||_2^2, \quad \forall w, y$$

$$\Downarrow \; y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Expected Bounded Stochastic Gradients

$$\mathbb{E}_j[||\nabla f_j(w^t)||_2^2] \leq B^2, \text{ for all iterates } w^t \text{ of SGD}$$

Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2} ||y - w||_2^2, \quad \forall w, y$$

$$y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Expected Bounded Stochastic Gradients

$$\mathbb{E}_j[||\nabla f_j(w^t)||_2^2] \leq B^2, \text{ for all iterates } w^t \text{ of SGD}$$

Assumptions for Convergence

Strong Convexity

$$f(y) \geq f(w) + \langle \nabla f(w), y - w \rangle + \frac{\lambda}{2} ||y - w||_2^2, \quad \forall w, y$$

$$y = w^*$$

$$2\langle \nabla f(w), w - w^* \rangle \geq \lambda ||w - w^*||_2^2$$

Expected Bounded Stochastic Gradients

$$\mathbb{E}_j[||\nabla f_j(w^t)||_2^2] \leq B^2, \text{ for all iterates } w^t \text{ of SGD}$$