SDCA for Regularized Loss Minimization

Shai and Tong

2013 年 6 月 5 日

Xinkai

Main Contritbution

Main Contritbution

Proving the convergence rate of duality gap for SDCA.

Main Contritbution

Proving the convergence rate of duality gap for SDCA.

Unfortunately

Main Contritbution

Proving the convergence rate of duality gap for SDCA.

Unfortunately

We'll skip the proofs.

Idea

Recall SGD

$$\omega := \omega - c \bigtriangledown P_i(\omega),$$

where our problem is: $\min P(\omega)$

Then how about SDCA?

Idea

Recall SGD

$$\omega := \omega - c \bigtriangledown P_i(\omega),$$

where our problem is: $\min P(\omega)$

Then how about SDCA?

Idea

Recall SGD

$$\omega := \omega - c \bigtriangledown P_i(\omega),$$

where our problem is: $\min P(\omega)$

Then how about SDCA?

Idea

Recall SGD

$$\omega := \omega - c \bigtriangledown P_i(\omega),$$

where our problem is: $\min P(\omega)$

Then how about SDCA?

DCA First

DCA First

Go from "$\min P(\omega)$" to "$\max D(\alpha)$"

DCA First

Go from "$\min P(\omega)$" to "$\max D(\alpha)$"

Instead of d primal variables $(\omega_1, \cdots, \omega_d)$, we have n dual variables $(\alpha_1, \cdots, \alpha_n)$,

DCA First

Go from "$\min P(\omega)$" to "$\max D(\alpha)$"

Instead of d primal variables $(\omega_1, \cdots, \omega_d)$, we have n dual variables $(\alpha_1, \cdots, \alpha_n)$, where d is the number of features and n is the number of entries.

DCA First

Go from "$\min P(\omega)$" to "$\max D(\alpha)$"

Instead of d primal variables $(\omega_1, \cdots, \omega_d)$, we have n dual variables $(\alpha_1, \cdots, \alpha_n)$, where d is the number of features and n is the number of entries.

Maximize dual in coordinate direction

DCA First

Go from "$\min P(\omega)$" to "$\max D(\alpha)$"

Instead of d primal variables $(\omega_1, \cdots, \omega_d)$, we have n dual variables $(\alpha_1, \cdots, \alpha_n)$, where d is the number of features and n is the number of entries.

Maximize dual in coordinate direction

For direction α_i, we find $\Delta\alpha_i$ that maximizes $D_i(\alpha_i^{t-1} + \Delta\alpha_i)$, where $D(\alpha) = \sum_i D_i(\alpha)$, and t is the loop index.

DCA First

Go from "$\min P(\omega)$" to "$\max D(\alpha)$"

Instead of d primal variables $(\omega_1, \cdots, \omega_d)$, we have n dual variables $(\alpha_1, \cdots, \alpha_n)$, where d is the number of features and n is the number of entries.

Maximize dual in coordinate direction

For direction α_i, we find $\Delta\alpha_i$ that maximizes $D_i(\alpha_i^{t-1} + \Delta\alpha_i)$, where $D(\alpha) = \sum_i D_i(\alpha)$, and t is the loop index. So $\lfloor \frac{t}{n} \rfloor$ is the number of epoch.

Then let's "S" (dca)

Random without Repetition (permutation)

Random with Repetition

Cyclic

Then let's "S"(dca)

Random without Repetition (permutation)

Random with Repetition

Cyclic

That's all

Then let's "S"(dca)

Random without Repetition (permutation)

Random with Repetition

Cyclic

That's all for the idea.

Then let's "S"(dca)

Random without Repetition (permutation)

Random with Repetition

Cyclic

That's all for the idea.

Questions?

Exp.0: Effect of different “S”

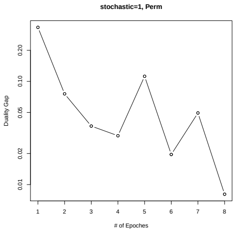


Figure: Perm

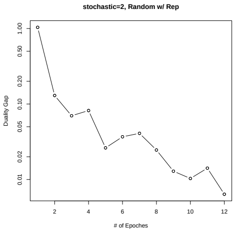


Figure: R w Rep

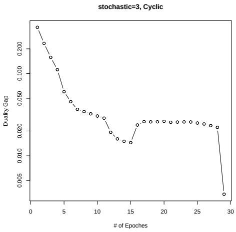


Figure: Cyclic

Exp.1: Convergence Rate of non-smooth Hinge

Result

For non-smooth hinge loss (we saw in class), the duality gap of SDCA should converge sub-linearly ($O(n + \frac{1}{\lambda\epsilon})$).

Exp.1: Convergence Rate of non-smooth Hinge

Result

For non-smooth hinge loss (we saw in class), the duality gap of SDCA should converge sub-linearly ($O(n + \frac{1}{\lambda\epsilon})$).

Experiment

Fix n and ϵ, vary $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$.

Exp.1: Convergence Rate of non-smooth Hinge

Result

For non-smooth hinge loss (we saw in class), the duality gap of SDCA should converge sub-linearly ($O(n + \frac{1}{\lambda\epsilon})$).

Experiment

Fix n and ϵ, vary $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$.

We would expect $\cdots$

Conti Exp.1: Convergence Rate of non-smooth Hinge

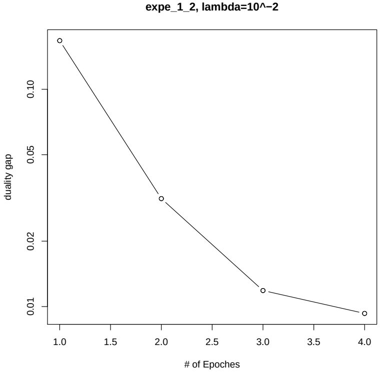


Figure: $\lambda = 10^{-2}$

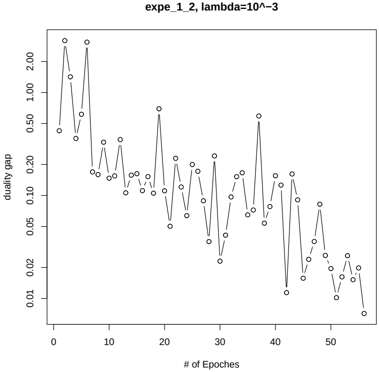


Figure: $\lambda = 10^{-3}$

Conti Exp.1: Convergence Rate of non-smooth Hinge

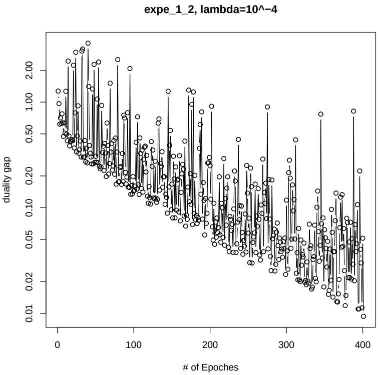


Figure: $\lambda = 10^{-4}$

Conti Exp.1: Convergence Rate of non-smooth Hinge

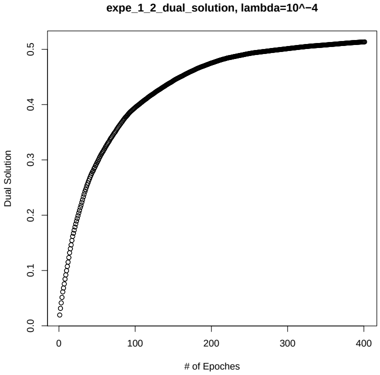


Figure: Dual Sol

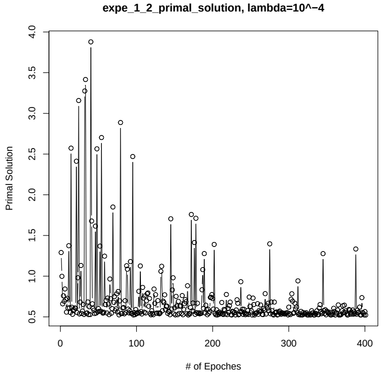


Figure: Primal Sol

Exp.2: Convergence Rate

Result

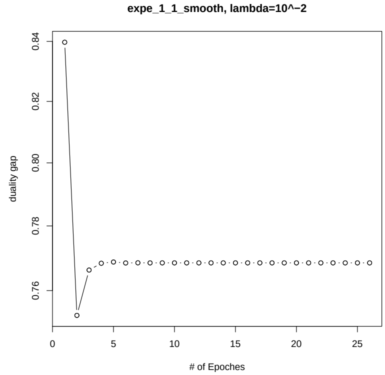
For smooth hinge loss, the duality gap of SDCA should converge linearly.

Have a look at code.

Exp.2: Convergence Rate

Result

For smooth hinge loss, the duality gap of SDCA should converge linearly.



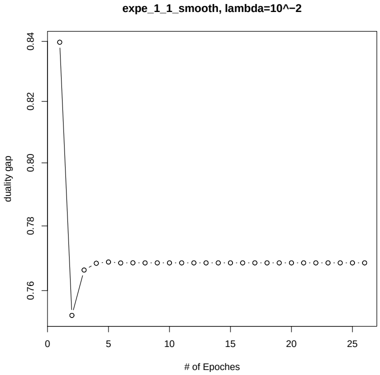
Duality Gap for Smooth Hinge.

Have a look at code.

Exp.2: Convergence Rate

Result

For smooth hinge loss, the duality gap of SDCA should converge linearly.



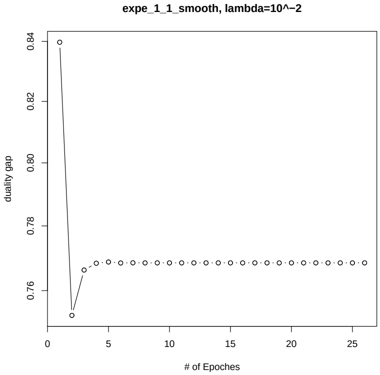
Duality Gap for Smooth Hinge.

`Have a look at code.`

Exp.2: Convergence Rate

Result

For smooth hinge loss, the duality gap of SDCA should converge linearly.

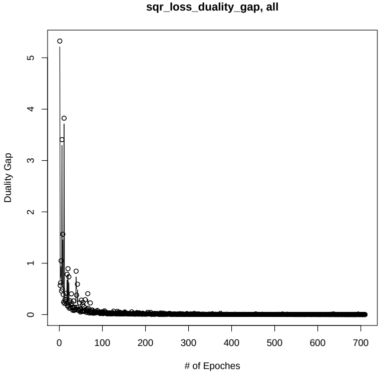


Duality Gap for Smooth Hinge.

`Have a look at code.`

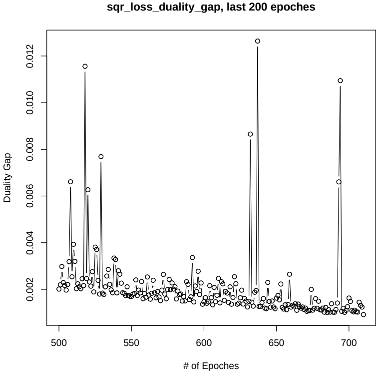
Exp.3: Convergence Rate of squared loss

Exp.3: Convergence Rate of squared loss



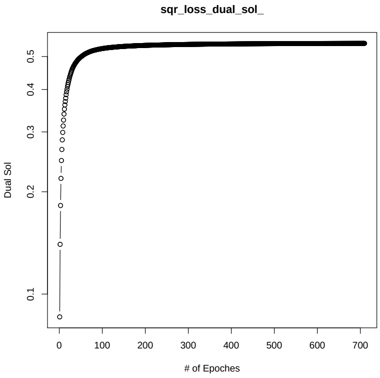
Duality Gap for Squared Loss

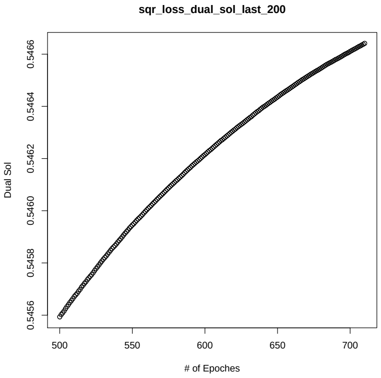
Conti Exp.3: Convergence Rate of squared loss



Telescoping the last 200 epoches

Conti Exp.3: Convergence Rate of squared loss, Dual Sol





Conti Exp.3: Convergence Rate of squared loss, Primal Sol