Coordinate Optimization Methods for Machine Learning

Nikita Doikov

Higher School of Economics

Seminar on Machine Learning, Voronovo, Russia
April 20, 2018

Plan of the talk

1. Empirical Risk Minimization problem

2. Coordinate Gradient Descent

3. Combining all together

Plan of the talk

1. Empirical Risk Minimization problem
2. Coordinate Gradient Descent
3. Combining all together

Empirical Risk Minimization

ERM problem:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \equiv \sum_{i=1}^{n} \underbrace{\phi_i(b_i^T w)}_{\text{loss}} + \underbrace{g(w)}_{\text{regularizer}} \right]$$

- $\{b_1, \ldots, b_n\}$ — given data, $b_i \in \mathbb{R}^d$ — features of i-th object.
- $w \in \mathbb{R}^d$ — weights of the model,
- $\{\phi_1, \ldots, \phi_n\}$ and g — convex functions:

$$\phi_i(\alpha x + (1-\alpha) y) \le \alpha \phi_i(x) + (1-\alpha)\phi_i(y), \quad \forall x, y \in \mathbb{R}, \forall \alpha \in [0, 1].$$

ERM problem:

$$\min_{w\in\mathbb{R}^d}\left[\; P(w) \;\equiv\; \sum_{i=1}^{n}\underbrace{\phi_i(b_i^T w)}_{\text{loss}} + \underbrace{g(w)}_{\text{regularizer}} \;\right]$$

- SVM: $\phi_i(a) = \max\{0, 1 - y_i a\}$,
- Logistic regression: $\phi_i(a) = \log(1 + \exp(-y_i a))$,
- Regression: $\phi_i(a) = (a - y_i)^2$ or $\phi_i(a) = |a - y_i|$,
- Support vector regression: $\phi_i(a) = \max\{0, |a - y_i| - \nu\}$,
- Generalized linear models.

Regularizers: $g(w) = \|w\|_2^2$, $g(w) = \|w\|_1$, indicator of simple set.

Why convexity?

- Local minimum = global minimum.
- Among only efficiently-solvable continuous problems.
- You can do a lot with convex models.
- Convex problems are used as subproblems in nonconvex optimization.
- Advanced methods from convex optimization work empirically well in nonconvex cases.

Dual Problem

$$
\begin{aligned}
\min_{w\in\mathbb{R}^d} P(w) &= \min_{w\in\mathbb{R}^d} \left[\sum_{i=1}^{n} \phi_i(\underbrace{b_i^T w}_{\equiv\mu_i}) + g(w)\right] \\
&= \min_{\substack{w\in\mathbb{R}^d \\ \mu\in\mathbb{R}^n \\ b_i^T w=\mu_i}} \left[\sum_{i=1}^{n} \phi_i(\mu_i) + g(w)\right] \\
&= \min_{\substack{w\in\mathbb{R}^d \\ \mu\in\mathbb{R}^n}} \max_{\alpha\in\mathbb{R}^n} \left[\sum_{i=1}^{n} \phi_i(\mu_i) + g(w) + \sum_{i=1}^{n} \alpha_i(b_i^T w - \mu_i)\right] \\
&\geq \max_{\alpha\in\mathbb{R}^n} \min_{\substack{w\in\mathbb{R}^d \\ \mu\in\mathbb{R}^n}} \left[\sum_{i=1}^{n} \phi_i(\mu_i) + g(w) + \sum_{i=1}^{n} \alpha_i(b_i^T w - \mu_i)\right] \\
&\equiv \max_{\alpha\in\mathbb{R}^n} D(\alpha)
\end{aligned}
$$

Fenchel Conjugate

Define: $\quad f^*(s) \equiv \sup_x \left[s^T x - f(x) \right].$

Then, we can rewrite objective of the Dual problem:

$$D(\alpha) \equiv \min_{\substack{w \in \mathbb{R}^d \\ \mu \in \mathbb{R}^n}} \left[\sum_{i=1}^n \phi_i(\mu_i) + g(w) + \sum_{i=1}^n \alpha_i (b_i^T w - \mu_i) \right]$$

$$= \sum_{i=1}^n -\phi_i^*(\alpha_i) - g^*\left(-B^T \alpha \right),$$

where rows of $B \in \mathbb{R}^{n \times d}$ are b_i^T.

We know conjugates of many functions!

Primal and Dual ERM problems

Primal problem:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \equiv \sum_{i=1}^{n} \phi_i(b_i^T w) + g(w) \right]$$

Dual problem:

$$\max_{\alpha \in \mathbb{R}^n} \left[D(\alpha) \equiv \sum_{i=1}^{n} -\phi_i^*(\alpha_i) - g^*\left(-B^T \alpha\right) \right]$$

- We know: $\min_{w \in \mathbb{R}^d} P(w) \geq \max_{\alpha \in \mathbb{R}^n} D(\alpha)$ — weak duality.
- Under very mild assumptions: $\min_{w \in \mathbb{R}^d} P(w) = \max_{\alpha \in \mathbb{R}^n} D(\alpha)$ — strong duality.
- There is a link between variables: $w = \nabla g^*\left(-B^T \alpha\right)$.

Example: Logistic regression

Primal problem:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \equiv \sum_{i=1}^{n} \log(1 + \exp(b_i^T w)) + \frac{\lambda}{2} \|w\|^2 \right]$$

- $\phi_i(a) = \log(1 + \exp(a)) \quad \Rightarrow$

  $\phi_i^*(s) = s \ln s + (1-s) \ln(1-s)$ for $s \in [0, 1]$.
- $g(w) = \frac{\lambda}{2} \|w\|^2 \quad \Rightarrow \quad g^*(v) = \frac{1}{2\lambda} \|v\|^2$

Dual problem:

$$\max_{\substack{\alpha \in \mathbb{R}^n, \\ 0 \le \alpha_i \le 1}} \left[D(\alpha) \equiv \sum_{i=1}^{n} -\alpha_i \ln \alpha_i - (1-\alpha_i) \ln(1-\alpha_i) - \frac{1}{2\lambda} \left\| -B^T \alpha \right\|^2 \right]$$

Example: SVM

Primal problem:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \equiv \sum_{i=1}^{n} \max\{0, 1 - y_i \cdot b_i^T w\} + \frac{\lambda}{2}\|w\|^2 \right]$$

- $\phi_i(a) = \max\{0, 1 - y_i \cdot a\} \quad \Rightarrow \quad \phi_i^*(s) = y_i s + \delta_{[0,1]}(-y_i s)$
- $g(w) = \frac{\lambda}{2}\|w\|^2 \quad \Rightarrow \quad g^*(v) = \frac{1}{2\lambda}\|v\|^2$

Dual problem:

$$\max_{\substack{\alpha \in \mathbb{R}^n, \\ 0 \leq -y_i \alpha_i \leq 1}} \left[D(\alpha) \equiv \sum_{i=1}^{n} -y_i \alpha_i - \frac{1}{2\lambda}\left\| -B^T \alpha \right\|^2 \right]$$

Example: LASSO

Primal problem:

$$\min_{w \in \mathbb{R}^d} \left[P(w) \equiv \frac{1}{2} \sum_{i=1}^{n} (y_i - b_i^T w)^2 + \lambda \|w\|_1 \right]$$

- $\phi_i(a) = \frac{1}{2}(y_i - a)^2 \quad \Rightarrow \quad \phi_i^*(s) = \frac{1}{2}s^2 + y_i \cdot s$
- $g(w) = \lambda \|w\|_1 \quad \Rightarrow \quad g^*(v) = \begin{cases} 0, & \|v\|_\infty \le \lambda, \\ +\infty, & \text{else} \end{cases}$

Dual problem:

$$\max_{\substack{\alpha \in \mathbb{R}^n, \\ \left\| -B^T \alpha \right\|_\infty \le \lambda}} \left[D(\alpha) \equiv \sum_{i=1}^{n} -\frac{1}{2}\alpha_i^2 - y_i \cdot \alpha_i \right]$$

Primal problem vs. Dual problem

$$\min_{w\in\mathbb{R}^d}\left[\sum_{i=1}^{n}\phi_i(b_i^T w)+g(w)\right] \;=\; \max_{\alpha\in\mathbb{R}^n}\left[\sum_{i=1}^{n}-\phi_i^*(\alpha_i)-g^*\left(-B^T\alpha\right)\right]$$

- **Big n.** Work in the **primal space**.
  — Process one loss function (= one example) at a time
  — Type of methods: stochastic gradient descent (SGD, SAG, SVRG, S2GD, SAGA, MISO, FINITO, ...)
- **Big d.** Work in the **primal space**.
  — Process one primal variable at a time
  — Type of methods: coordinate gradient descent.
- **Big n.** Work in the **dual space**.
  — Progress one dual variable (= one example) at a time.
  — Type of methods: coordinate gradient descent.

Werner Fenchel, 1905 – 1988

Werner Fenchel, 1972

Plan of the talk

1. Empirical Risk Minimization problem
2. Coordinate Gradient Descent
3. Combining all together

Problem:

$$\min_{x \in \mathbb{R}^n} f(x),$$

f is convex and differentiable, n is Huge.

- **Gradient Descent:**

$$x^{k+1} = x^k - \alpha_k \cdot \nabla f(x^k).$$

- **Coordinate Gradient Descent:**

$$x^{k+1} = x^k - \alpha_k \cdot (\nabla f(x^k))_{i_k} \cdot e_{i_k},$$

$i_k \in \{1, \dots, n\}$ – selected coordinate,
e_{i_k} – basis vector:

$$e_{i_k} \equiv (0, 0, \dots, 0, 0, \underbrace{1}_{i_k}, 0, 0, \dots, 0, 0)^T \in \mathbb{R}^n.$$

Initialization: choose $x^0 \in \mathbb{R}^n$.

Iteration $k \geq 0$:

- Pick $S_k \subseteq \{1, \ldots, n\}$.
- Do a step:
$$x^{k+1} = x^k - \alpha_k \cdot [\nabla f(x^k)]_{S_k}$$

Notation: $$[\nabla f(x^k)]_{S_k} \equiv \sum_{i \in S_k} (\nabla f(x^k))_i \cdot e_i.$$

Note: cost of k-th iteration is proportional to $|S_k|$.

Main issues

- How to choose coordinates?

  — Randomly!

- How to choose α_k?

- How to generalize method?

  — Constrains, nondifferentiable components?

- Rate of convergence?

Key observation

- Functions with Lipschitz continuous gradients:

$$\|\nabla f(x) - \nabla f(y)\| \;\leq\; L\|x-y\|, \quad \forall x, y \in \mathbb{R}^n.$$

  Then, we have a global upper bound:

$$f(y) \;\leq\; f(x) + \langle \nabla f(x), y-x \rangle + \frac{L}{2}\|y-x\|^2, \quad \forall x, y \in \mathbb{R}^n.$$

- Gradient Step: $\quad x^{k+1} \;=\; x^k - \frac{1}{L}\nabla f(x_k) \;=$

$$= \operatorname*{argmin}_{y\in\mathbb{R}^n} \Big[f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{L}{2}\|y-x^k\|^2 \Big]$$

- Coordinate Gradient Step: $\quad x^{k+1} \;=\; x^k - \frac{1}{L}[\nabla f(x^k)]_{S_k} \;=$

$$= \operatorname*{argmin}_{\substack{y\in\mathbb{R}^n, \\ y_i = x_i^k \text{ for } i \notin S_k}} \Big[f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{L}{2}\|y-x^k\|^2 \Big].$$

Coordinate Gradient Step

$$x^{k+1} \;=\; x^k \;-\; \alpha_k \cdot [\nabla f(x^k)]_{S_k} \;=$$

$$= \quad \mathop{\mathrm{argmin}}_{\substack{y \in \mathbb{R}^n, \\ y_i = x_i^k \text{ for } i \notin S_k}} \quad \Bigl[f(x^k) + \langle \nabla f(x^k), y - x^k \rangle + \frac{1}{2\alpha_k} \| y - x^k \|^2 \Bigr],$$

Thus, step of the method is just a minimization of regularized linear model of f on the random subspace. $S_k \subseteq \{1, \dots, n\}$.

- How to choose α_k? **Key inequality:**

$$f(x^{k+1}) \;\leq\; f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{1}{2\alpha_k} \| x^{k+1} - x^k \|^2 \quad \textbf{(*)}$$

— It is enough to set $\alpha_k \equiv \frac{1}{L}$.

— **Do an adaptive search:** decrease twice α_k until **(*)** holds; increase twice every iteration.

Generalization of the method

From unconstrained to constrained optimization:

$$\min_{x\in\mathbb{R}^n} f(x) \qquad \longrightarrow \qquad \min_{x\in Q} f(x),\ Q\subseteq\mathbb{R}^n.$$

- *Basic* method for $\min_{x\in\mathbb{R}^n} f(x)$:

$$x^{k+1} = \operatorname*{argmin}_{\substack{y\in\mathbb{R}^n,\\ y_i=x_i^k \text{ for } i\notin S_k}} \left[f(x^k) + \langle\nabla f(x^k), y-x^k\rangle + \frac{1}{2\alpha_k}\|y-x^k\|^2\right],$$

- Method for $\min_{x\in Q} f(x)$:

$$x^{k+1} = \operatorname*{argmin}_{\substack{\color{red}{y\in Q},\\ y_i=x_i^k \text{ for } i\notin S_k}} \left[f(x^k) + \langle\nabla f(x^k), y-x^k\rangle + \frac{1}{2\alpha_k}\|y-x^k\|^2\right],$$

Rate of convergence

- f is convex, with Lipschitz-continuous gradient:

$$\|\nabla f(x) - \nabla f(y)\| \;\leq\; L\|x - y\|, \quad \forall x, y \in \mathbb{R}^n.$$

- Step size α_k satisfies **key inequality (*)**. (For example, $\alpha_k \equiv \frac{1}{L}$)
- Uniform sampling:

$$\mathbb{P}(i \in S_k) = \mathbb{P}(j \in S_k) \equiv p, \quad \forall i, j \in \{1, \dots, n\}.$$

$$\tau \;\equiv\; \mathbb{E}\big[|S_k|\big] \;=\; \mathbb{E}\textstyle\sum_{i=1}^n [i \in S_k] \;=\; \sum_{i=1}^n \mathbb{E}[i \in S_k] \;=\; np.$$

Theorem

$$\mathbb{E}\big[f(x^k)\big] - f^* \quad \leq \quad \frac{2}{k} \cdot \frac{n}{\tau} \cdot \max\big\{LD^2, f(x^0) - f^*\big\},$$

$D \stackrel{\text{def}}{=} \sup\big\{\|x - x^*\| \mid f(x) \leq f(x^0)\big\}$

Iteration complexity

In order to get

$$\mathbb{E}\left[f(x^K)\right] - f^* \;\leq\; \varepsilon,$$

it is enough to set

$$K \;=\; \frac{2}{\varepsilon} \cdot \frac{n}{\tau} \cdot \max\left\{LD^2, f(x^0) - f^*\right\}.$$

- $O\left(\frac{1}{\varepsilon}\right)$ iterations. Compare with SGD: $O\left(\frac{1}{\varepsilon^2}\right)$.
- Remind: $\tau \equiv \mathbb{E}\left[|S_k|\right]$. Cost of one iteration: $O(\tau)$.
- Lipschitz constant L often proportional to data matrix.
- Logisitic regression and SVM: $L = \|B\|$.

Plan of the talk

1. Empirical Risk Minimization problem
2. Coordinate Gradient Descent
3. Combining all together

- We want to solve

ERM problem:

$$\min_{w \in \mathbb{R}^d} \Big[\, P(w) \;\equiv\; \sum_{i=1}^{n} \phi_i(b_i^T w) \;+\; g(w) \, \Big]$$

- We construct

Dual ERM problem:

$$\max_{\alpha \in \mathbb{R}^n} \Big[\, D(\alpha) \;\equiv\; \sum_{i=1}^{n} -\phi_i^*(\alpha_i) \;-\; g^*\Big(-B^T \alpha\Big) \, \Big]$$

- Apply Coordinate Gradient method to this formulation!
- Why we can not apply it to the primal?

Coordinate Gradient Ascent for SVM

Dual SVM problem:

$$\max_{\substack{\alpha \in \mathbb{R}^n, \\ 0 \le -y_i \alpha_i \le 1}} \left[D(\alpha) \equiv \sum_{i=1}^{n} -y_i \alpha_i - \frac{1}{2\lambda} \left\| -B^T \alpha \right\|^2 \right]$$

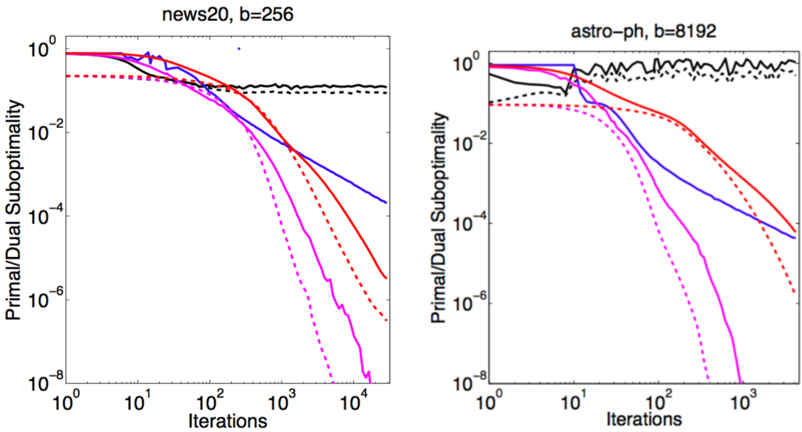
Label of i-th object: $y_i \in \{-1, 1\}$.

- Choose **one** random coordinate per step ($\tau \equiv \mathbb{E}|S_k| = 1$).
- $(\nabla D(\alpha))_i = -y_i - \frac{1}{\lambda}(BB^T \alpha)_i$.
- One method step: $\tilde{\alpha}_i^{k+1} = \alpha_i^k - c_k \cdot (\nabla D(\alpha))_i$.
- Projection onto $0 \le -y_i \alpha_i \le 1$:

$$\alpha_i^{k+1} = \begin{cases} 0, & \text{if} \quad -y_i \tilde{\alpha}_i^{k+1} < 0, \\ -y_i, & \text{if} \quad -y_i \tilde{\alpha}_i^{k+1} > 1, \\ \tilde{\alpha}_i^{k+1}, & \text{else.} \end{cases}$$

- Recompute primal variables: $w_{k+1} = -\frac{1}{\lambda} B^T \alpha_{k+1}$.

Experiments: training SVM.



Legend: **SGD** – **Blue**, **SDCA** – **Purple** .

Source: *Mini-Batch Primal and Dual Methods for SVMs, 2013, Takac et al.*

The Big Picture

- **Stochastic Gradient Descent (SGD):**
  - \+ Strong theoretical guarantees.
  - – Hard to tune step size (requires $\alpha \to 0$).
  - – No clear stopping criterion.
  - – Converges fast at first, then slow to more accurate solution.
- **Coordinate Gradient Ascent for the Dual problem:**
  - \+ Strong theoretical guarantees (better than SGD).
  - \+ Easy to tune step size (adaptive line search).
  - \+ Terminate when the duality gap is sufficiently small.
  - \+ Converges to accurate solution faster than SGD.